现代数学手册

·经济数学卷

Modern Mathematics Handbook

《现代数学手册》编纂委员会

· 华中科技大学出版社 ·

现代数学手册

MODERN MATHEMATICS HANDBOOK

经济数学卷

《现代数学手册》编纂委员会





• 华中科技大学出版社 •
(华中理工大学出版社)
中国·武汉

《现代数学手册》编纂委员会

顾　　问　钱伟长　吴文俊　杨叔子

主　　编　徐利治

副 主 编　张尧庭　林化夷　卢开澄

分卷主编　经典数学卷　廖晓昕

　　　　　近代数学卷　胡适耕

　　　　　计算机数学卷　卢开澄

　　　　　随机数学卷　陈希孺　郑忠国

　　　　　经济数学卷　王国俊　施光燕

　　　　　(以下按姓氏笔画为序)

编　　委　王兴华　王能超　毛经中　叶其孝

　　　　　史树中　李国伟　苏维宜　余家荣

　　　　　佘健棠　陈文忠　周蕴时　胡毓达

执行编委　佘健棠　林化夷　郭永康　姜新祺

责任编辑　龙纯曼　叶见欣　李立鹏　佟文珍

　　　　　佘健棠　周芬娜　姜新祺

前　言

在人类开始跨入21世纪的历史时期，人们已普遍地看到了一种历史现象，即数学问题的多样性与数学应用的广泛性及深入性，已经成为现代科技发展的重要特征。可以预期，伴随着计算机科技在新世纪里的不断发展，此特征今后还将以更高的水平显示出来。

在中国，“科学技术是第一生产力”（邓小平名言）已逐渐成为人们信奉的朴实真理。国家富强显然要以第一生产力即科技的发达为必要条件。但是，如果没有近、现代发展起来的数学各分支学科作工具，当然也就不会有现代科技。因此“国家富强必须要依靠数学发达”这句经典名言（拿破仑（Napoleon）名言），自然也是一条不容置疑的客观真理。

基于上述认识，在华中理工大学出版社的倡议与委托下，我们通过集体协作，努力编纂了这部《现代数学手册》巨著，其目的正是怀着对我国将在新世纪里能尽快成为富强国家的热切希望，而欲为科技界提供一份力所能及的奉献。具体说来，这部工具性巨著服务的读者（或使用者）对象，包括广大科学工作者、工程技术人员、经济管理工作者、高等院校的教师和学生等。

那么，作为数学工具书，这部巨型手册要求具备哪些特点呢？在编写过程中，出版社负责人和我们达成了一项共识，即手册应具备科学性、先进性、实用性、规范性与简明性。200余位撰稿人与审稿人（来自中国科学院、北京大学、清华大学、复旦大学、南京大学、浙江大学、北京师范大学、厦门大学、上海交通大学、西安交通大学、中国科技大学、南开大学、武汉大学、华中理工大学、大连理工大学、南京航空航天大学、陕西师范大学等40多所高校与研究所）按照这些特点和要求付出了

艰辛的劳动。我们要感谢他们的通力合作与努力,使本手册基本上体现了上述所希冀的特点或特色。

为了读者选购和使用方便,本手册分5卷出版,分别名为"经典数学卷"、"近代数学卷"、"计算机数学卷"、"随机数学卷"和"经济数学卷"。需要指出的是,各个分支(篇目)的归属是相对的,这里考虑了各分卷篇幅大小的平衡问题。例如,"蒙特卡罗法"这一篇也可归入"计算机数学卷"。

我们要感谢诸分卷主编为精心组稿、编稿、审稿付出的精力和时间。特别要对中国科学院两位老院士钱伟长先生与吴文俊先生,以及杨叔子院士乐愿担任本手册的顾问而致以诚挚的谢忱。最后,还要对华中理工大学出版社具有远见卓识的负责人和埋头苦干的编辑人员与我们在本手册的生产全过程中的互相配合和精诚合作,深表谢忱。

《现代数学手册》编纂委员会

主编 徐利治

1999年12月于武汉

现代数学手册

篇目录

经典数学卷

近代数学卷

计算机数学卷

随机数学卷

经济数学卷

MODERN MATHEMATICS HANDBOOK

CONTENTS

CLASSICAL MATHEMATICS

MODERN MATHEMATICS

COMPUTER MATHEMATICS

STOCHASTIC MATHEMATICS

ECONOMIC MATHEMATICS

目　录

·经济数学卷·

第1篇

计量经济

编　者　林少宫
审校者　冯文权

目　录

引　言

计量经济学(econometrics)是20世纪30年代开始发展起来的一门实证性数量经济学科.它运用数理统计和统计推断的方法,对由经济理论指引的经济关系(式)进行实测和经济研究,为经济预测和政策的制定提供依据.

计量经济学可大致分为理论计量经济学和应用计量经济学.前者主要研究适合于测算计量经济模型所设定的经济关系式的统计方法及其性质(如广泛使用的最小二乘法及其性质);后者则考虑如何运用前者去研究经济学中的某些特殊领域,如生产函数、消费函数、投资函数或货币需求函数,等等.

计量经济学从方法和步骤上可划分为四个部分:模型设定、参数估计、验证、预测和控制.即首先建立某一经济理论的数学模型,用数学语言特别是用方程式表达经济理论;其次通过参数估计把方程式中的参数值填人;然后按一定的准则验证所依据的经济理论是否可以接受,有时需要将模型作适当修改,重新估计参数,再验证模型;最后,利用所估算的模型进行预测和控制.

计量经济学方法所面对的数据主要是非实验(nonexperimental)数据①.这一特点有助于说明它为什么不仅是近代经济学、金融学研究中不可缺少的技术,而且在社会、人文、历史以至管理工程等行为科学方面都有广泛的应用.

1　线性回归模型

线性模型　它是最简单而又最常用的经济模型或经济关系式.经济模型可以是线性或非线性的."线性(一次)"(linearity)既可对变量而言,也可对参数而言.在计量经济学中,作为一种方便易行的方法或形式而采用的线性模型,是指它所含的每个方程对参数而言都是线性的.例如,边际成本方程

$$Y=\alpha+\beta X+\gamma X^2,$$

对变量(产量)X虽然是非线性的,但对参数α,β和γ是线性的,所以被认为是线性的.方程

$$Y=\alpha+\sqrt{\beta}X,$$

对变量X是线性的而对参数β不是,所以被认为是非线性的.相对于参数而言的线性模型之所以被广泛应用,主要因为它便于对参数作最小二乘(least squares)估计,而且许多原来不是线性的函数形式,经过一定的变换,就可变为线性.例如柯布-道格拉斯(Cobb-Douglas)生产函数

---

① 由于现代实验经济学的发展,这句话只有相对意义.

$$Q = AK^{\alpha}L^{\beta},$$

其中 K, L 分别表示资本、劳力投入, Q 表示产出,取其对数就变成对参数 α 和 β 来说为线性函数.

在现实中,经济关系(式)不是一种确切的关系,可在线性关系式中适当加入一个误差(干扰或扰动)项 u,如

$$Y = \alpha + \beta X + u, \tag{1-1}$$

其中,u 表示在考虑 Y 和 X 的线性关系时被忽略掉的许多微小因素的总和.由于方程(1-1)中已设有截距项 α,就不妨考虑 u 的均值或数学期望为零,即

$$E(u) = 0.$$

这里 E 是数学期望符号.于是,$\alpha + \beta X$ 就成为在给定 X 时 Y 的条件均值或条件数学期望.因为,将(1-1)式两边求数学期望,得

$$E(Y|X) = E(\alpha + \beta X + u) = \alpha + \beta X + E(u),$$

即

$$E(Y|X) = \alpha + \beta X. \tag{1-2}$$

方程(1-1)表示 Y 对 X 的**回归方程**.回归的目的是要求出形如方程(1-2)的条件期望值.为此,不但要估计其中的参数 α 和 β,而且,为了知道估计的好坏,一般还要估计误差 u 的方差,记为 σ_u^2,

$$\sigma_u^2 = E(u^2).$$

在回归方程中,作为可以给定的自变量 X,又通称为**回归元**(regressor),而随之而变的应变量 Y,则通称为**回归值**(regressand).在计量经济学中,因应用目的的不同,Y 和 X 曾被赋予种种不同的名称,如表 1-1 所列举.

表 1-1

Y		X	
应变量	(dependent variable)	自变量	(independent variable)
被解释变量	(explained v)	解释变量	(explanatory v)
预测值	(predictand)	预测元	(predictor)
回归值	(regressand)	回归元	(regressor)
内生变量	(endogenous v)	外生变量	(exogenous v)
响应值	(response)	刺激量	(stimuli)
目标值	(target)	控制变量	(control v)
		(政策变量)	(policy v)

1.1 二元线性回归

含 k 个自变量的线性回归方程可写为

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u. \tag{1-3}$$

当 $k=2$ 时为二元线性回归方程;当 $k>2$ 时为多元线性回归方程.为了估计方程中

的参数 β_i,须对 Y 和诸 X 进行 $n(n>k)$次观测.观测数据可分横截面和时间序列两种情形.当数据来自横截面时,常记观测结果为

$$Y_i=\beta_1+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+u_i,\quad i=1,2,\cdots,n.$$

而当数据来自时间数列时,则记为

$$Y_t=\beta_1+\beta_2X_{2t}+\beta_3X_{3t}+\cdots+\beta_kX_{kt}+u_t,\quad t=1,2,\cdots,n(\text{或 } T).$$

自然,u_i 或 u_t 是不可观测的.当含义明显或无须区分横截面或时间序列时,可略去下标 i 或 t 而直接写为(1-3)式.

1.1.1 最小二乘法与经典假设

最小二乘法 是计量经济学中最基本和最常用的参数估计方法.以二元线性回归方程

$$Y_i=\alpha+\beta X_i+u_i,\quad i=1,2,\cdots,n$$

为例.现要作出估计:

$$\hat{Y}_i=a+bX_i.$$

其中,a,b 分别是α,β 的估计值.最小二乘法要求估计值 a,b 使得对 Y_i 的估计误差平方和 $\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2$ 最小.令 $e_i=\hat{Y}_i-Y_i$,代表最小二乘(回归)残差(或剩余),则要求

$$\sum_{i=1}^{n}e_i^2=\min\sum_{i=1}^{n}(Y_i-(a+bX_i))^2.$$

对 a,b 求偏导数并令其为零,得所谓正规方程(normal equations)

$$\frac{\partial(\sum e_i^2)}{\partial a}=-2\sum(Y_i-a-bX_i)=0\quad(\text{即}\sum e_i=0);$$

$$\frac{\partial(\sum e_i^2)}{\partial b}=-2\sum X_i(Y_i-a-bX_i)=0\quad(\text{即}\sum X_ie_i=0).$$

记

$$x_i=X_i-\bar{X},\quad \bar{X}=\sum X_i/n;$$

$$y_i=Y_i-\bar{Y},\quad \bar{Y}=\sum Y_i/n.$$

解正规方程得

$$b=\frac{\sum x_iy_i}{\sum x_i^2},\quad a=\bar{Y}-b\bar{X},$$

这就是 β 和α 的最小二乘(LS)估计.

平方和分解 它是对线性回归实行最小二乘法后的一种直观分析.容易推出

$$\sum(Y_i-\bar{Y})^2=\sum(\hat{Y}_i-\bar{Y})^2+\sum(Y_i-\hat{Y}_i)^2,$$

记作

$$\underset{(\text{总平方和})}{\text{TSS}} = \underset{(\text{解释平方和})}{\text{ESS}} + \underset{(\text{剩余平方和})}{\text{RSS}}$$

从而把 Y 的总平方和分解为由 X 解释了的解释平方和以及 X 未能解释的剩余(残差)平方和两部分.

定义

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}.$$

r^2(或 r^2_{YX})表示总平方和 TSS 中由 X 解释了或者说由回归解释了的部分,最小二乘法将使这个部分达到最大. $r = \pm\sqrt{r^2}$称为相关系数. r 取与 b 相同的符号.

经典假设 u_i 的期望值为零,方差为 σ^2(不因 i 而异),u_i 和 $u_j(i \neq j)$不相关,且 u_i 与 X_i 无关,即

$$E(u_i) = 0;\quad E(u_iu_j) = \begin{cases} 0 & (i \neq j), \\ \sigma^2 & (i = j); \end{cases}\quad \text{cov}(u_i, X_i) = 0. \tag{1-4}$$

其中,E 是数学期望符号,cov 是协方差符号.可得到 σ^2(或 σ^2_u)的最小二乘估计

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}. \tag{1-5}$$

还可以证明这个估计是无偏的,即 $E(\hat{\sigma}^2) = \sigma^2$.

上述关于误差项 u_i 的假定常被喻为经典假设.

高斯-马尔可夫定理 在 u_i 满足经典假设的条件下,参数(α,β)的最小二乘估计(a,b)为**最优线性无偏估计**(best linear unbiased estimate,简记 BLUE).最优是指估计量的抽样方差达到最小,线性是对 Y_i 而言,无偏是指估计量的期望值等于被估参数,而且 $\hat{Y}_i = a + bX_i$ 也是 Y_i 的最优线性无偏估计.

1.1.2 正态性假设与 t 统计量

在上述关于最小二乘法的经典假设(1-4)式的基础上,现再假定 u_i 服从正态分布

$$u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2),$$

其中 $N(0,\sigma^2)$表示均值为零、方差为 σ^2 的正态分布,i.i.d.表示独立同分布,即每次观测误差 u_i 都是独立的并有相同的正态概率分布.这样,将最小二乘估计 b 标准化为 Z,就得到

$$Z = \frac{b - E(b)}{\sigma_b} = \frac{b - \beta}{\sigma_b} \sim N(0,1),$$

即标准化的 b 服从标准正态分布,其中 σ_b 表示 b 的标准误(差).可以证明①

① 在 u 的经典假设下,这个方差是在 β 的线性无偏估计中最小的;如再加上 u 的正态性假设,则是所有无偏估计中的最小方差.

$$\sigma_b^2 = \frac{\sigma_u^2}{\sum x_i^2}.$$

当 σ_u^2 未知而代以它的最小二乘无偏估计 $\hat{\sigma}^2$(见(1-5)式)时，Z 变量就变为服从自由度为 $(n-2)$ 的 t 分布，

$$t = \frac{b-\beta}{\hat{\sigma}_b} \sim t_{n-2}.$$

其中 $\hat{\sigma}_b^2 = \frac{\sum e_i^2}{n-2} \frac{1}{\sum x_i^2}$. 这样，可通过 t 分布对 β 做统计推断.

相仿，可以证明 a 的方差

$$\sigma_a^2 = \frac{\sigma^2 \sum X_i^2}{n \sum x_i^2},$$

若 σ^2 未知，则代之以 $\hat{\sigma}^2$，并记

$$\hat{\sigma}_a^2 = \frac{\hat{\sigma}^2 \sum X_i^2}{n \sum x_i^2}.$$

同样可得

$$t = \frac{a-\alpha}{\hat{\sigma}_a} \sim t_{n-2},$$

从而又可通过 t 分布对 α 做统计推断.

例 1 现代投资分析中所谓特征线(characteristic line)就是来自如下模型的一条回归线：

$$r_{it} = \alpha_i + \beta_i r_{mt} + u_t.$$

其中，
r_{it} = 证券 i 在时刻 t 的回报率；
r_{mt} = 市场证券组合在时刻 t 的回报率；
u_t = 随机扰动项.

模型中的 β_i，称为证券 i 的 β 系数，代表一种证券的市场(或系统)风险. 现根据 1956～1976 年的 240 个月回报率算出相对于某研究机构所开发的市场证券指数的某公司股票特征线：

$$r_{it} = 0.7264 + 1.0598 r_{mt}, \quad r^2 = 0.4710,$$

$$\hat{\sigma}_{a_i} = 0.3001, \quad \hat{\sigma}_{b_i} = 0.0728, \quad f = 240 - 2 = 238.$$

其中 f 为 t 分布的自由度. 由金融理论知，当 β_i 系数大于 1 时，证券 i 属于进攻型；当截距 $\alpha_i \neq 0$ 时，证券 i 不是均衡定价的. 因此检验如下假设是有明确的经济意义的：

(1) $H_0: \beta_i \leqslant 1, H_1: \beta_i > 1$ (单边假设)；

(2) $H_0: \alpha_i = 0, H_1: \alpha_i \neq 0$ (双边假设).

在假设 H_0 属真的情况下，可算出：

对 α_i 有

$$t=\frac{a_i-\alpha_i}{\widehat{\sigma}_{a_i}}=\frac{0.7264-0}{0.3001}=2.4213;$$

对 β_i 有

$$t=\frac{b_i-\beta_i}{\widehat{\sigma}_{b_i}}=\frac{1.0598-1}{0.0728}=0.8214.$$

因为自由度 238 很大，t 分布无异于正态分布，查表知 a_i 显著异于零而拒绝 $H_0:\alpha_i=0$；但 b 并不显著大于 1，因此接受 $H_0:\beta\leqslant 1$.

本例通过显著性的计算判断了所得数值结果的统计意义. 除此之外，在计量经济中，还要讲求这些数值结果的经济意义. 比如说，喜欢冒更多风险的人会乐意于 $\beta_i>1$ 的股票买卖. 因为每当市场回报率增加时，他得到的回报率将增加更多. 当然，当市场回报率减少时他的回报率也将减少更多. 同理，比较保守的人就愿意购买 $\beta_i<1$ 的股票. 另外，根据资本性资产定价模型(CAPM)原理，证券 i 的平均回报正比于风险 β_i，理应 $\alpha_i=0$. 当 $\alpha_i>0$ 时，表示回报高于此比例，由此推知证券 i 的市场定价过低；反之，$\alpha_i<0$ 时表示定价过高. 因此，所作假设 $H_0:\alpha=0$ 和 $H_0:\beta_i\leqslant 1$ 是有明确的经济意义的. 在计量经济中，检验所估算的模型的统计显著性固然重要，待检验的假设本身的经济含义(economic significance)也同样重要.

1.1.3　参数估计的最大似然法

对回归模型 $Y_i=\alpha+\beta X_i+u_i,\ i=1,2,\cdots,n$ 来说，假定

$$u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2),$$

就是假定

$$Y_i \overset{\text{i.i.d.}}{\sim} N(\alpha+\beta X_i,\sigma^2).$$

记正态密度函数为 ϕ，则有

$$\phi(Y_i)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{(Y_i-(\alpha+\beta X_i))^2}{\sigma^2}\right\}.$$

由于 n 次观测的独立性，$Y_i(i=1,2,\cdots,n)$的联合密度函数为

$$\begin{aligned}\phi(Y_1,Y_2,\cdots,Y_n)&=\phi(Y_1)\ \phi(Y_2)\cdots\phi(Y_n)\\&=\frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left\{-\frac{1}{2}\sum\frac{(Y_i-\alpha-\beta X_i)^2}{\sigma^2}\right\}.\end{aligned}$$

在此密度函数中，X_i,α,β 及 σ^2 被看作已知，故上式可写为

$$\phi(Y_1,Y_2,\cdots,Y_n)=\phi(Y_1,Y_2,\cdots,Y_n\mid X_i;\alpha,\beta,\sigma^2).$$

若反过来把 $Y_1,\cdots,Y_n$ 看作已知或给定，而 α,β 及 σ^2 被看作未知，则 ϕ 称为 α，β 及 σ^2 的似然函数(likelihood function)，且记为

$$L(\alpha,\beta,\sigma^2\mid X_i;Y_i)=\frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left\{-\frac{1}{2}\sum\frac{(Y_i-\alpha-\beta X_i)^2}{\sigma^2}\right\},$$

或取其对数

$$\ln L = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum(Y_i - \alpha - \beta X_i)^2.$$

参数 α,β 及 σ^2 的**最大似然估计法**(method of maximum likelihood)就是要解决当 X_i, $Y_i(i=1,2,\cdots,n)$给定时,α,β,σ^2 取何值能使似然函数(或其对数 $\ln L$)达到最大的问题.

容易看出,欲使 L(或 $\ln L$)最大,须使

$$\sum(Y_i - \alpha - \beta X_i)^2$$

最小.可见 α,β 的最大似然(ML)估计就是它们的最小二乘(LS)估计.

但是,对 σ^2 的 ML 估计 $\tilde{\sigma}^2$,有

$$\tilde{\sigma}^2 = \frac{\sum\hat{u}_i^2}{n}.$$

(这里的 $\hat{u}_i$ 即本篇 1.1.1 节中的 $e_i = \hat{Y}_i - Y_i$,代表 LS 回归残差)它不等于 LS 估计 $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\sum\hat{u}_i^2}{n-2}.$$

因此,ML 估计 $\tilde{\sigma}^2$ 是有偏误的.但随着 n 的增大,$\tilde{\sigma}^2$ 将是渐近无偏的.

理论上 ML 估计有许多优良的大样本性质,如渐近有效性、渐近正态性、一致性、不变性等.

1.1.4 时序相关性

在误差的经典假设(1-4)式中,曾假定

$$E(u_iu_j) = \begin{cases} 0, & i \neq j; \\ \sigma^2, & i = j. \end{cases}$$

即误差是同方差的且无序列相关.然而,在现实中,许多经济关系式中的误差或者呈异方差性,或者有序列相关性.

时序相关(自相关)(serial correlation(auto-correlation)) 它是指一个变量的现在值和它的过去值的相关.在平稳时间序列的分析中,应着重考虑的不是经济变量本身的自相关,而是误差项的自相关.出现误差自相关有以下多种原因:

(1)经济行为中的因循守旧习惯,如消费习惯.

(2)模型的设定误差,如在需求函数中,忽略替代品价格;在边际成本函数中忽略产量的二次效应;在供给函数中忽略多期滞后现象,等等.

(3)数据修匀后呈现规则性.

(4)在横截面中相邻单位的仿效作用、聚类现象,等等.

误差序列和它的滞后序列的相关关系,称自相关或序列相关(不排除有序号的横截面数据也会有序列(序号)相关),其一般模型为

$$Y_t = \alpha + \beta X_t + u_t, \quad t = 1,2,\cdots.$$

对 $t \neq s$,

$$\mathrm{cov}(u_t, u_s) = E(u_t u_s) \neq 0.$$

协方差不为零表明了误差 u_t 的时序相关.

常用的一阶自相关模型为

$$\begin{aligned} &Y_t = \alpha + \beta X_t + u_t, \quad t = 1, 2, \cdots, T; \\ &u_t = \rho u_{t-1} + \varepsilon_t; \\ &E(\varepsilon_t) = 0; \quad E(\varepsilon_t \varepsilon_s) = \begin{cases} 0 & (t \neq s), \\ \sigma_\varepsilon^2 & (t = s); \end{cases} \quad |\rho| < 1. \end{aligned} \tag{1-6}$$

其中,(1-6)式表明 u_t 遵从一阶自回归过程(即 u_t 为一阶自回归序列). 可以求出,对一切 t,

$$\begin{aligned} &E(u_t) = 0, \\ &E(u_t^2) = \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1 - \rho^2}, \\ &E(u_t u_{t-s}) = \rho^s \sigma_u^2, \end{aligned} \tag{1-7}$$

因而

$$\rho = \frac{E(u_t u_{t-1})}{\sigma_u^2}$$

为 u_t 与 u_{t-1}之间的相关系数. (1-7)式给出过程的自相关函数 ρ^s 为 $E(u_t u_{t-s})/\sigma_u^2$,其图形称为**相关图**(correlogram).

忽略自相关而用普通最小二乘(OLS)法作回归估计时,往往严重地歪曲了回归系数 β 的真像. β 的普通最小二乘估计量,记为 b_{OLS},会有很大的波动,而且按普通最小二乘法计算的 b_{OLS}的抽样方差 $\mathrm{var}(b_{OLS})$,当 X 有正的序列相关且 $\rho > 0$ 时,可能大大地低估了它的真值. 但是,仍然可以证明,估计量 b_{OLS}是无偏的.

德宾-沃森(D-W)检验是最常用以检验一阶自相关的方法,所用的检验统计量称 **d 统计量**,其定义为

$$d = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2},$$

其中 $\hat{u}_t$ 是最小二乘回归误差(残差). 容易推知,

$$d \approx 2\left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right) = 2(1 - \hat{\rho}), \tag{1-8}$$

其中 $\hat{\rho}$代表一阶自相关模型中的相关系数 ρ 的估计值. 由于$|\rho| \leqslant 1$,知 d 围绕2而波动,其波动范围从零到 4. 当且仅当 $\hat{\rho} > 0$ 时, $d < 2$;当且仅当 $\hat{\rho} < 0$ 时, $d > 2$.

因 d 的抽样分布依赖于数据 X 的结构,故德宾-沃森只能建立 d 的相对于某个显著水平的上界 d_U 和下界 d_L. 先考虑 $d < 2$ 的一侧. 对应于一定的显著水平(5%或 1%),可查表(D-W 统计量 d 检验表)得到 d_L 和 d_U.

若 $d < d_L$,则认为有正的自相关;

若 $d > d_U$,则认为无自相关;

若 $d_L \leqslant d \leqslant d_U$,则不作结论.

至于负相关,则取对称于 2 的另一侧,即 $d > 2$ 的一侧:

若 $4 - d_L < d < 4$,则认为有负的自相关;

若 $d < 4 - d_U$,则认为无自相关;

若 $4 - d_U \leqslant d \leqslant 4 - d_L$,则不作结论.

参看下面的 d 检验示意图 1-1.

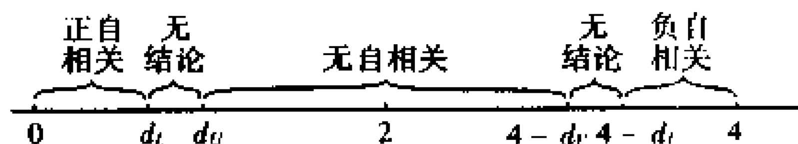


图 1-1 d 检验示意图

d_L 和 d_U 的确定仅与观测次数 T、解释变量的个数 k 以及显著水平有关. 注意,d 检验表的计算假定了回归方程含有截距项并且 X 是非随机的,因此,当滞后的内生变量作为解释变量出现时,d 检验表便不适用了.

由于时序相关对回归估计有较严重的影响,现在在回归估算程序中,除了计算回归系数的抽样方差(或标准误)和回归方程的相关系数外,通常都伴随有 D-W 统计量 d 的计算结果.

为了解决在一阶自相关情形下的回归估计问题,可采用**广义最小二乘法**,设法通过变量转换,把不满足经典假设的误差项变为能满足的情形,然后实行普通最小二乘法. 其步骤如下:

把一阶自回归模型中的回归方程滞后一期,乘以(1-6)式中的相关系数 ρ,然后,从原回归方程中减去. 即

$$Y_t = \alpha + \beta X_t + u_t,$$

$$Y_{t-1} = \alpha + \beta X_{t-1} + u_{t-1},$$

$$(Y_t - \rho Y_{t-1}) = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \varepsilon_t. \tag{1-9}$$

或记

$$Y_t^* = \alpha^* + \beta^* X_t^* + \varepsilon_t, \quad \varepsilon_t = u_t - \rho u_{t-1}. \tag{1-10}$$

其中,

$$Y_t^* = Y_t - \rho Y_{t-1}, \quad X_t^* = X_t - \rho X_{t-1}, \quad \alpha^* = \alpha(1-\rho), \quad \beta^* = \beta.$$

Y_t^* 和 X_t^* 就是 Y_t 和 X_t 的转换变量. 由于 ε_t 满足经典假设,故适合于用 OLS 法估计(1-10)中的参数 α^* 和 β^*,然后通过转换关系

$$\alpha^* = \alpha(1-\rho) \quad 和 \quad \beta^* = \beta$$

解出 α 和 β. 当然,问题还在于怎样求解 ρ.

求解 ρ 的方法,一种方法是利用(1-8)式中的估计值 $\hat{\rho}$. 另一种方法是采用**柯克伦-奥克特(Cochrane-Orcutt)迭代法**,把(1-9)式写成另一形式,即

$$(Y_t - \alpha - \beta X_t) = \rho(Y_{t-1} - \alpha - \beta X_{t-1}) + \varepsilon_t. \tag{1-11}$$

给定一个 ρ 值,便可从(1-9)式得到 α,β 的 LS 估计,且记为 $\hat{\alpha},\hat{\beta}$. 将 $\hat{\alpha},\hat{\beta}$代入(1-11)式后,又可利用(1-11)式得到 ρ 的 LS 估计,且记为 $\hat{\rho}$. 再将 $\hat{\rho}$代入(1-9)式又可得 α,β 的第二次 LS 估计,且记为 $\hat{\hat{\alpha}},\hat{\hat{\beta}}$. 再代入(1-11)式以估计 ρ. 如此反复迭代,直至相继两次估计值的差异落在设定的范围内. 一般只要迭代3次至4次便足够好.

关于(一阶)自相关模型的估计问题还有许多方法以及伴随而来的一些困扰问题,如是否漏掉某些自变量,等等,这里就不一一叙述了.

1.1.5　异方差性

回归模型中,误差项的异方差性是回归估计所面临的另一重要问题. **异方差性**(heteroscedasticity)是指在样本的第 i 次观测中误差 u_i 的方差 σ_i^2 随 i 而异,结合回归模型可表示为

$$Y_i = \alpha + \beta X_i + u_i,$$
$$E(u_i)=0,\quad E(u_i^2)=\sigma_i^2,\quad E(u_iu_j)=0,\quad 当\ i\neq j.$$

异方差性 $E(u_i^2)=\sigma_i^2$ 是相对于**同方差性**(homoscedasticity) $E(u_i^2)=\sigma^2$ 而言的. 异方差性的经济现象是很多的. 例如,一个家庭的收入越多,它的储蓄波动就越大. 参看图1-2和图1-3. 一个企业的规模越大,工人所得报酬可能越大,但报酬的波动也越大. 在一个学习的回归模型中,将会发现学习的时间越长,出现差错的方差越小. 一般地说,在横截面回归模型中比在时间序列模型中异方差性的问题更为普遍.

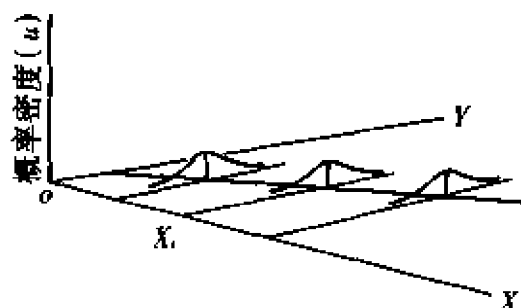


图1-2　同方差性

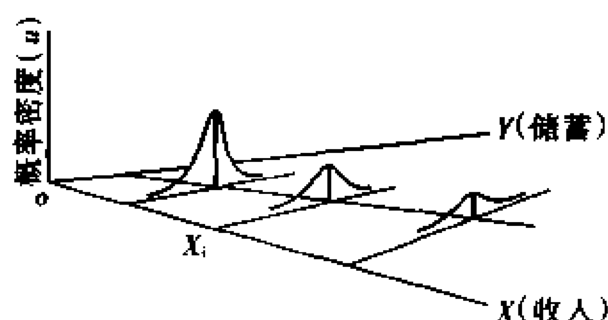


图1-3　异方差性

当异方差出现时,普通最小二乘法虽然仍给出线性无偏估计,但这个估计将不是最小方差的. 可以利用加权或广义最小二乘法,以期得到接近于最优的线性无偏估计. 假定知道 $\sigma_i^2, i=1,2,\cdots,n$. 将第 i 次观测的回归方程通除以 σ_i,有

$$\frac{Y_i}{\sigma_i}=\frac{\alpha}{\sigma_i}+\beta\frac{X_i}{\sigma_i}+\frac{u_i}{\sigma_i},$$

将 u_i 转换为 u_i/σ_i,后者自然是同方差的,其方差为1,从而 LS 估计是 BLUE. 当然,这里又有一个问题是:怎样知道或怎样估计 σ_i? 在实践中,往往考虑 σ_i 同 X_i(或 Y_i)的某个幂函数有比例关系而用该幂函数去代替 σ_i,或者利用时间序列回归提供的残差去估计 σ_i.

在有异方差情形下，回归系数 β 的普通最小二乘估计 $\widehat{\beta}_{OLS}$ 的方差将不是 $\frac{\sigma^2}{\sum x_i^2}$ 而是 $\frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$．若取 $\mathrm{var}(\widehat{\beta}_{OLS})=\sigma^2/\sum x_i^2$，则 $\widehat{\beta}_{OLS}$ 既可能偏大，也可能偏小，要看 X_i 和 σ_i 之间有怎样的相关关系．但可以证明，对 β 的广义最小二乘估计 $\widehat{\beta}_{GLS}$，必有

$$\mathrm{var}(\widehat{\beta}_{GLS})\leqslant \mathrm{var}(\widehat{\beta}_{OLS}).$$

使用 $\mathrm{var}(\widehat{\beta}_{OLS})$ 进行统计推断，所造成后果的严重性，要看异方差程度的大小而定．对于横截面回归，常有必要检验每一具体应用中的异方差性问题．有关文献中提供了许多检验方法，然而没有哪一种方法能被推荐为较有通用价值的标准方法．所有方法的基本思想都是通过对原回归方程的 OLS 估计，看其 LS 残差平方 $\widehat{u}_i^2$ 或绝对值 $|\widehat{u}_i|$ 是否与某一或某些变量（特别是原回归方程中的 X 自变量）有某种明显的相关．有时为了能较彻底地检验异方差性是否显著，要通过多种检验作出判断．

1.1.6 自回归条件异方差性

自回归条件异方差性（auto regressive conditional heteroscedasticity 简记为 ARCH）的发现，是近代金融统计研究的一个重要成果，表明在时间序列的分析中，不仅要考虑序列（特别是误差序列）本身的自相关，还可考虑其平方的自相关．例如人们发现股票价格变化的一个明显特点，往往是大（小）的波动尾随着大（小）的波动．因此，股价的方差不仅随前时刻的波动情况而异，也就是在时间上有条件异方差性，而且在时序上还呈现正的（自）相关．这样的一些实际问题引发了对 ARCH 模型的大量研究及其在金融市场上的广泛应用．异方差性的分析和应用并不限于横截面回归．

ARCH 的一个简单表述方法是通过以下的一个误差模型来表示：

$$u_t=\epsilon_t(\lambda_0+\lambda_1u_{t-1}^2)^{1/2}\quad(\lambda_0>0,0\leqslant\lambda_1<1),\tag{1-12}$$

其中，$\epsilon_t\sim N(0,1)$，$E(\epsilon_t\epsilon_s)=0$，当 $t\neq s$；$E(\epsilon_tu_s)=0$，当 $t>s$．这时，将能导出 u_t 的无条件方差

$$\begin{aligned}\mathrm{var}(u_t)&=E(u_t^2)=E(\epsilon_t^2)E(\lambda_0+\lambda_1u_{t-1}^2)\\&=\lambda_0+\lambda_1E(u_{t-1}^2)=\lambda_0+\lambda_1\lambda_0+\lambda_1E(u_{t-2}^2)\\&=\cdots=\lambda_0(1+\lambda_1+\lambda_1^2+\cdots)=\frac{\lambda_0}{1-\lambda_1}\end{aligned}$$

为一常数．然而，对给定的 u_{t-1}，u_t 的**条件方差**却是

$$\begin{aligned}\mathrm{var}(u_t|u_{t-1})&=E(\epsilon_t^2|u_{t-1})E(\lambda_0+\lambda_1u_{t-1}^2|u_{t-1})\\&=\lambda_0+\lambda_1u_{t-1}^2,\end{aligned}$$

这一条件方差随前一时刻的波动 u_{t-1}^2 而变化．当然，还可以在模型(1-12)的右边引入更多的滞后项．在应用中，误差 u_t 往往联系着一个回归函数，比如，

$$Y_t = \alpha + \beta X_t + u_t, \quad t = 1, \cdots, n,$$

u_t 是 Y_t 对 X_t 的回归方程中的误差项.

在有 ARCH 型异方差性情形下,由于回归的误差项毕竟是无条件同方差性的,而且不难看出它又是没有自相关的(但不能说是独立的),因此最小二乘回归估计仍然是 BLUE.但如考虑非线性的最大似然估计,则可以获得更好的估计.为此,可用$(\lambda_0 + \lambda_1 u_{t-1}^2)^{1/2}$通除回归方程,得

$$Y_t(\lambda_0 + \lambda_1 u_{t-1}^2)^{-1/2} = \alpha(\lambda_0 + \lambda_1 u_{t-1}^2)^{-1/2} + \beta X_t(\lambda_0 + \lambda_1 u_{t-1}^2)^{-1/2} + \epsilon_t,$$

以消除误差项的条件异方差性并使其独立.为了得到 α, β, λ_0 和 λ_1 的似然函数,通过变量代换得

$$f(Y_t) = \left|\frac{\mathrm{d}\epsilon_t}{\mathrm{d}Y_t}\right| f(\epsilon_t) = (\lambda_0 + \lambda_1 u_{t-1}^2)^{-1/2} f(\epsilon_t)$$

及

$$\ln f(Y_1, Y_2, \cdots, Y_n) = -\frac{1}{2}\sum \ln(\lambda_0 + \lambda_1 u_{t-1}^2) + \ln f(\epsilon_1, \epsilon_2, \cdots, \epsilon_n).$$

于是得到所求对数似然函数(注意 $\mathrm{var}(\epsilon_t) = 1$):

$$\ln L = -\frac{1}{2}\sum_{t=1}^{n}\ln(\lambda_0 + \lambda_1(Y_{t-1} - \alpha - \beta X_{t-1})^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left\{\frac{(Y_t - \alpha - \beta X_t)^2}{\lambda_0 + \lambda_1(Y_{t-1} - \alpha - \beta X_{t-1})^2}\right\}.$$

可见 ML 估计是非线性的.

1.2 多元线性回归

1.2.1 多元线性回归与最小二乘法

类似于二元线性回归,对多元线性回归模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \cdots, n$$

中的参数 β_i 作最小二乘(LS)估计,是指求 $\widehat{\beta}_i$ 使得

$$Q = \min\sum_{i=1}^{n}(Y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 X_{2i} + \cdots + \beta_k X_{ki}))^2.$$

$\widehat{\beta}_i$ 可从一组对$\widehat{\beta}_i$ 为线性的所谓正规方程

$$\frac{\partial Q}{\partial \beta_i} = 0, \quad i = 1, \cdots, k \tag{1-13}$$

中解出,从而得到样本回归方程

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_{2i} + \cdots + \widehat{\beta}_k X_{ki}.$$

同样,可将 Y_i 的总平方和(TSS)分解为由 X_i 或由回归解释的解释平方和(ESS)及剩余平方和(RSS),

$$\underset{\text{TSS}}{\sum(Y_i-\bar{Y})^2}=\underset{\text{ESS}}{\sum(\hat{Y}_i-\bar{Y})^2}+\underset{\text{RSS}}{\sum(Y_i-\hat{Y}_i)^2},$$

并类似地定义多元回归的**判定系数** R^2 为

$$R^2=\frac{\text{ESS}}{\text{TSS}}.$$

R 称为 Y 对全部 X_i, $i=2,3,\cdots,k$ 的**复相关系数**. 显然, $|R|\leqslant 1$. 由于自变量不止一个, Y 与 X 的关系是正是负不明,故通常计算 R^2.

在 u_i 满足经典假设的条件下, LS 估计 $\hat{\beta}_i$ 仍然是 BLUE.

为了保证正规方程组对 β_i 有唯一解,要求诸 X_i, $i=1,2,\cdots,k$ 无共线性,即不存在一组非零的 $\lambda_1,\lambda_2,\cdots,\lambda_k$,使得

$$\lambda_1X_1+\lambda_2X_2+\cdots+\lambda_kX_k=0,$$

其中 $X_1\equiv 1$. 也就是说,诸 X_i 是线性无关的. 反之, X_i 就是线性相关的. 在线性相关的情形下,任一 X_j 可以表示为其余 X_i 的线性组合. 例如

$$X_2=-\frac{\lambda_1}{\lambda_2}X_1-\frac{\lambda_3}{\lambda_2}X_3-\cdots-\frac{\lambda_k}{\lambda_2}X_k,$$

这表明 X_2 与其他 $X_i(i\neq 2)$ 有完全的线性相关,其相关系数为 1. 实际上,至少是由于观测上的误差,不会出现上述情形. 真实的问题却是存在有一组非 0 的 $\lambda_1,\lambda_2,\cdots,\lambda_k$,使得

$$\lambda_1X_1+\lambda_2X_2+\cdots+\lambda_kX_k\approx 0. \tag{1-14}$$

这个接近于零的等式表明某 X_j 与其余 $X_i(i\neq j)$ 有高度相关, **多重共线性**(multicollinearity)指的就是由(1-14)式引起的估计问题. 无碍于一般性,假定 $k=3$,由最小二乘法(即由正规方程(1-13)),有

$$\hat{\beta}_2=\frac{(\sum y_ix_{2i})(\sum x_{3i}^2)-(\sum y_ix_{3i})(\sum x_{2i}x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2)-(\sum x_{2i}x_{3i})^2},$$

$$\text{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum x_{2i}^2(1-r_{23}^2)}.$$

其中 r_{23} 表示 X_2 与 X_3 的相关系数. 如果 $r_{23}^2=\frac{(\sum x_2x_3)^2}{(\sum x_2^2)(\sum x_3^2)}=1$, 则 $\hat{\beta}_2$ 无解. 而且,当 $r^2\to 1$ 时, $\text{var}(\hat{\beta}_2)\to\infty$, 即 $\hat{\beta}_2$ 作为 β_2 的估计是非常不可靠的.

在二元回归方程 $Y=\alpha+\beta X+u$ 中,称 β 为**回归系数**;在多元回归

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_kX_k+u \tag{1-15}$$

中,称 β_i, $i=2,3,\cdots,k$, 为**偏回归系数**. 虽然当 $k=2$ 时, β_2 无异于二元回归中的回归系数 β,但除非 X_2 与其余的 X_i, $i=3,4,\cdots,k$ 不相关,作为二元回归($k=2$)来估计 β_2 和作为多元回归($k>2$)来估计 β_2,估计的结果是不一样的.

1.2.2 偏回归系数与偏相关系数

下面解释偏回归系数的含义如何不同于 $k=2$ 时的回归系数. 为了便于解释,

下面把二元回归方程写成

$$Y = \beta_{11} + \beta_{12} X_2 + u. \tag{1-16}$$

在多元回归方程(1-14)中,偏回归系数 β_2 的含义是,在保持 $X_3, X_4, \cdots, X_k$ 不变的条件下,X_2 每变化1单位,平均而言 Y 将变化 β_2 单位,或者说条件(数学)期望 $E(Y|X_3, X_4, \cdots, X_k)$ 将变化 β_2 单位.但实际上,当 X_2 变化时,$X_3, \cdots, X_k$ 会随之而变.所谓"保持"不变,是"统计意义上"的不变.无碍于一般性,且考虑 $k=3$ 的情形,有

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u,$$

及其最小二乘估计

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{u}. \tag{1-17}$$

"保持"X_3 不变,是指用统计方法去掉 X_3 在其变化中对 Y 和 X_2 的影响.通过 Y 对 X_3 的 LS 回归

$$Y = \hat{\beta}_{11} + \hat{\beta}_{13} X_3 + \hat{u}_1 \tag{1-18}$$

得到的残差 $\hat{u}_1$,就代表除掉 X_3 的影响之后的 Y.同理,再通过 X_2 对 X_3 的 LS 回归

$$X_2 = \hat{\beta}_{21} + \hat{\beta}_{23} X_3 + \hat{u}_2 \tag{1-19}$$

得到的残差 $\hat{u}_2$,就代表除去 X_3 的影响后的 X_2.

现在再求 $\hat{u}_1$ 对 $\hat{u}_2$ 的 LS 回归

$$\hat{u}_1 = \hat{\alpha}_1 + \hat{\alpha}_2 \hat{u}_2 + \hat{u}_3,$$

所得到的回归系数 $\hat{\alpha}_2$,便是上述三元回归方程(1-17)式中的偏回归系数 $\hat{\beta}_2$,即 $\hat{\alpha}_2 = \hat{\beta}_2$.

由此可见,(1-16)式中的回归系数 β_{12} 是没有去掉 X_3 的影响之前 Y 对 X_2 的回归系数(可称为简单回归系数),它含有 X_2 对 Y 的直接影响(β_2)和通过 X_3 对 Y 起的间接影响($\beta_3\beta_{32}$)两个部分.其最小二乘估计有如下关系:

$$\hat{\beta}_{12} = \hat{\beta}_2 + \hat{\beta}_3 \hat{\beta}_{32}.$$

其中,间接影响由 X_3 对 Y 的直接影响($\hat{\beta}_3$)乘以 X_2 对 X_3 的影响($\hat{\beta}_{32}$)而得到.易见,除非 X_2 与 X_3 不相关或 X_3 对 Y 无直接影响,简单回归系数将不等于偏回归系数.由于间接影响可正可负,$\hat{\beta}_{12}$ 既可大于也可小于 $\hat{\beta}_2$,两者甚至可以符号相反.这在解释多元回归时,应加注意.

偏回归系数的含义有一重要应用:如果怀疑 Y 和 X 之间的关系或相互影响是由第三者(比如说由时间 t 的变化)引起的,那么在 Y 对 X 的回归方程中增加一个代表第三者的自变量(比如 t),便能把问题弄清楚.

与偏回归有密切联系的一个概念是偏相关.仍以 $k=3$ 的三元回归为例,Y 与 X_2 的偏相关系数是指除去 X_3 的影响后的 Y 和 X_2 的相关系数,即 $\hat{u}_1$ 和 $\hat{u}_2$(见

(1-18)式和(1-19)式)的相关系数,记为 $r_{12.3}$,表示"固定"X_3 而得到的 Y 和 X_2 的相关系数,以区别于简单相关系数 $r=r_{12}$和复相关系数 $R=R_{1.23}$. 可以证明,对于 LS 回归,R^2,r_{12}^2和 $r_{12.3}^2$有如下关系:

$$R^2 = \underset{\text{(由 }X_2\text{ 解释的平方和)}}{r_{12}^2} + \underset{\text{(}X_2\text{ 未解释而由 }X_3\text{ 解释的平方和)}}{(1-r_{12}^2)r_{13.2}^2}.$$

上式可写成

$$(1-R^2)=(1-r_{12}^2)(1-r_{13.2}^2),$$

并容易推广到 k 元回归的情形:

$$1-R^2=(1-r_{12}^2)(1-r_{13.2}^2)(1-r_{14.23}^2)\cdots(1-r_{1k.23\cdots(k-1)}^2).$$

其中 $1-r_{1j.23\cdots(j-1)}^2$代表 $X_2,X_3,\cdots,X_{j-1}$所未能解释的平方和中 X_j 也未能解释的部分.

1.2.3 多元回归的一些特殊形式及其应用

1. 多项式回归

在多元线性回归模型

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+u$$

中,取 $X_i=X^i$,便是一个多项式回归,它在经济分析中有一定的应用场合. 例如根据理论分析,生产的边际成本先是递减然后递增. 记产量为 X,成本为 Y,关于边际成本的变化,就可用 X 的二次方程

$$Y=\beta_0+\beta_1X+\beta_2X^2+u$$

做 Y 的回归分析,而总成本的变化就可由三次多项式回归来表示. 低次多项式回归还适用于工资报酬与工作经验之间的描述,经济增长或发展(如国民产值在时间上的变化)的勾画,等等.

2. 虚拟变量(dummy variables)

在经济分析中常遇到一些定性变量,如性别、种族、季节等作为回归元(解释变量). 此时可把某些 X_i 虚拟为(0,1)变量. 例如

取 $X_1=\begin{cases}1 & \text{男性},\\ 0 & \text{女性};\end{cases}$

取 $X_1=\begin{cases}1 & \text{春季},\\ 0 & \text{非春季};\end{cases}\quad X_2=\begin{cases}1 & \text{夏季},\\ 0 & \text{非夏季};\end{cases}\quad X_3=\begin{cases}1 & \text{秋季},\\ 0 & \text{非秋季}.\end{cases}$

这里要注意,如果在回归方程中含有常数项,就不可再取 $X_4=\begin{cases}1 & \text{冬季}\\ 0 & \text{非冬季}\end{cases}$,以避免出现自变量 X_i 之间的完全共线性. 因为,不言而喻,已设的常数项即代表了冬季效应.

3. 滞后变量(lagged variables)

在经济关系中,一些回归元往往有一种持续效应. 例如收入对消费(或储蓄)一般都有较持久的影响. 消费不仅受当前收入并且受过去收入的影响. 这样,就可考虑含滞后项的回归,如

$$Y_t=\alpha+\beta_0X_t+\beta_1X_{t-1}+\cdots+\beta_kX_{t-k}+u_t.$$

其中 Y_t 代表时期 t 的消费，X_s 代表时期 s 的收入；α,β 为参数. 又如某农产品的供给 Y_t 依赖于一年前的价格 X_{t-1}，可将其回归模型写为

$$Y_t = \alpha + \beta X_{t-1} + \text{其他自变量的线性组合} + u_t.$$

如果使用季度数据，则可写

$$Y_t = \alpha + \beta X_{t-4} + \text{其他自变量的线性组合} + u_t.$$

在多元回归的应用中，不仅要考虑选择什么自变量作为解释变量，而且还要适当考虑这些自变量或解释变量的幂函数和滞后期，尽可能使模型接近于描述现实.

4. 对数变换

在线性回归中，无论对回归元 X 或对回归值 Y 取对数，对回归方程中的参数来说仍可以是线性的. 在经济分析中，根据具体问题，对 Y 或 X 适当作对数变换是有明确的经济含义的. 由于对数的变化是一种相对变化，或者说是百分比变化，对数 Y 相对于对数 X 的变化就是经济学中所描述的一种弹性. 例如，

$$\ln Y = \beta_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + u,$$

当 Y 代表需求，X_2 代表价格，X_3 代表收入时，β_2 和 β_3 就分别是需求的价格弹性和收入弹性. 又如，

$$\ln Y = \alpha + \beta X + u,$$

适用于描述国民收入 Y 与年度 X 之间的关系，其中

$$\beta = \text{国民收入的相对变化/时间(年)的绝对变化},$$

代表一定时期内国民经济的恒定增长率.

还可以考虑

$$Y = \alpha + \beta \ln X + u,$$

用于描述国民总产值 Y 与货币供给 X 的关系，等等.

1.2.4 F 检验统计量及其应用

计量经济学常借助于数理统计学中的假设检验手段来对计量模型进行验证和选择. 在多元回归中，F 统计量是检验假设的最重要也是最主要的检验统计量. 其方法是对模型中的参数作线性约束，借以比较未受约束的回归剩余平方和与受约束的回归剩余平方和，看两者的差异是否在"统计上"显著. 这里关键在于怎样把模型的验证或选择问题化为对参数的约束是否成立的问题.

在 LS 的理论背景假设下，增设误差项 u_i 的正态性，即

$$u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2),$$

则以下的 F 统计量（又称 F 比率）服从自由度为 f_1 和 f_2 的 F 分布，即

$$F = \frac{(\mathrm{RSS_R} - \mathrm{RSS_U})/f_1}{\mathrm{RSS_U}/f_2} \sim F_{f_1,f_2}, \tag{1-20}$$

其中，$\mathrm{RSS_R}$ 和 $\mathrm{RSS_U}$ 分别表示受约束和无约束的最小二乘回归剩余平方和，f_1 和 f_2 分别是分子和分母的自由度. 具体地说，f_1 = 约束的个数，f_2 = 误差（或剩余）平方和（$\mathrm{RSS_U}$）计算中的自由度.

例 2　对回归模型

$$Y_i=\beta_1+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+u_i,\quad i=1,2,\cdots,n,$$

检验(零)假设

$$H_0:\beta_2=\beta_3=\cdots=\beta_k=0,$$

即检验整个回归模型的显著性,对立假设为

$$H_1:\text{并非所有回归系数均为零}.$$

在假设 H_0 属真的情形下,(1-20)式可化为

$$F=\frac{R^2/(k-1)}{(1-R^2)/(n-k)},$$

其中 R 为复相关系数.此时 F 检验等价于 R^2 检验.对选定的显著水平 α,如 $F>F_{\alpha(k-1,n-k)}$,则拒绝 H_0,否则接受 H_0.

当 $k=2$ 时,$F=t^2$,F 检验就等价于 t 检验.

例 3　柯布-道格拉斯生产函数的回归模型

$$\ln Y_i=\beta_1+\beta_2\ln X_{2i}+\beta_3\ln X_{3i}+u_i,\quad i=1,2,\cdots,n,\tag{1-21}$$

其中 Y 代表产出,X_2 代表资本投入,X_3 代表劳力投入.现检验假设

$$H_0:\beta_2+\beta_3=1,$$
$$H_1:\beta_2+\beta_3\neq 1.$$

当约束 $\beta_2+\beta_3=1$ 成立时,

$$\ln Y=\beta_1+(1-\beta_3)\ln X_2+\beta_3\ln X_3+u_0$$

或

$$\ln(Y/X_2)=\beta_1+\beta_3\ln(X_3/X_2)+u_0.\tag{1-22}$$

其中,　(Y/X_2) = 产出/劳力投入比率,

(X_3/X_2) = 资本/劳力比率,

均有重要经济含义.用 LS 法分别对(1-21)式和(1-22)式计算回归剩余平方和,有

$$\sum\hat{u}^2=\mathrm{RSS_U}\quad\text{及}\quad\sum\hat{u}_0^2=\mathrm{RSS_R},$$

代入(1-20)式即可算出 $F_{(1,n-3)}$ 来同 $F_{\alpha(1,n-3)}$ 比较.当 $F>F_\alpha$ 时拒绝 H_0,否则接受 H_0.

例 4　**结构性变化的检验.**

设 Y = 储蓄,X = 个人收入.为判断两个时期的储蓄行为有无变化,考虑

第一个时期 Y 对 X 的回归

$$Y_t=\alpha_1+\alpha_2X_t+u_{1t},\quad t=1,2,\cdots,n_1;$$

第二个时期 Y 对 X 的回归

$$Y_t=\beta_1+\beta_2X_t+u_{2t},\quad t=1,2,\cdots,n_2;$$

并检验假设

$$H_0:\alpha_1=\beta_1\quad\text{且}\quad\alpha_2=\beta_2,$$
$$H_1:\alpha_1=\beta_1\quad\text{和}\quad\alpha_2=\beta_2\quad\text{不同时成立}.$$

为了应用 F 检验统计量,假定两时期的回归误差相互独立,将两时期的数据合并

使用以估计共同的回归方程

$$Y_t = \lambda_1 + \lambda_2 X_t + u_{3t}, \quad t = 1, 2, \cdots, (n_1 + n_2).$$

当 H_0 属真时,这样做是合理的.用LS法估计以上回归的剩余平方和

$$\sum \widehat{u_1}^2, \quad \sum \widehat{u_2}^2 \quad 及 \quad \sum \widehat{u_3}^2.$$

不难理解,无约束平方和将由

$$\mathrm{RSS_U} = \sum \widehat{u_1}^2 + \sum \widehat{u_2}^2$$

给出,而受约束平方和可由 $\sum \widehat{u_3}^2$ 来代表,即

$$\mathrm{RSS_R} = \sum \widehat{u_3^2}.$$

这样就算出

$$F = \frac{(\mathrm{RSS_R} - \mathrm{RSS_U})/2}{\mathrm{RSS_U}/((n_1 - 2) + (n_2 - 2))}$$

$$= \frac{(\sum \widehat{u_3}^2 - \sum \widehat{u_1}^2 - \sum \widehat{u_2}^2)/2}{(\sum \widehat{u_1}^2 + \sum \widehat{u_2}^2)/(n_1 + n_2 - 4)}.$$

当 $F > F_\alpha$ 时,以显著水平 α 拒绝 H_0,认为储蓄行为有所变化;否则认为储蓄行为无显著变化.

1.3 线性回归的矩阵表述

用矩阵表述线性回归的最大优点在于它的形式简洁,便于统一处理任意多个变元的回归模型.一旦把 k 个变元的回归模型建立并求解,其解式将适用于 k 等于2,3,以至任意多个变元的情形.但是,矩阵代数的简洁形式容易掩盖其内在的详细含义和微细差别,阅读时必须留意.

1.3.1 最小二乘回归的矩阵方法

一个 k 元线性回归模型可写为

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, 2, \cdots, n,$$

其中 n 代表对变元的观测次数,它是如下的一组 n 个方程的缩写:

$$Y_1 = \beta_1 + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + u_1,$$

$$Y_2 = \beta_1 + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + u_2,$$

$$\vdots$$

$$Y_n = \beta_1 + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + u_n.$$

用矩阵代数表示,便可写成

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

或编写成

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}} \underset{k\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{u}},$$

其中$\underset{n\times 1}{\boldsymbol{Y}}$表示 $\boldsymbol{Y}$ 为 n 行 1 列矩阵，$\underset{n\times k}{\boldsymbol{X}}$表示 $\boldsymbol{X}$ 为 n 行 k 列矩阵，其第一列可理解为$\underset{n\times 1}{\boldsymbol{X}_1}=1$的列元素，余类推．当行数和列数清楚而不致引起误解时，可仅写

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{u}. \tag{1-23}$$

用 $\boldsymbol{A}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{A}$ 的转置．关于误差项 $\boldsymbol{u}^{\mathrm{T}}=[u_1,u_2,\cdots,u_n]$的经典假设，可对照纯量情形列出表 1-2.

表 1-2　关于线性回归模型的经典假设

纯量表示	矩阵表示
1. $E(u_i)=0$，对每个 i	1. $E(\boldsymbol{u})=\boldsymbol{0}$，其中 $\boldsymbol{u}$ 和 $\boldsymbol{0}$ 均为 $n\times 1$ 列向量
2. $E(u_iu_j)=\begin{cases}0, & i\neq j;\\ \sigma^2, & i=j\end{cases}$	2. $E(\boldsymbol{uu}^{\mathrm{T}})=\sigma^2\boldsymbol{I}$，其中 $\boldsymbol{I}$ 为 $n\times n$ 恒等矩阵
3. $X_2,X_3,\cdots,X_k$ 为非随机或固定量	3. $n\times k$ 矩阵 $\boldsymbol{X}$ 为非随机的，即由一组固定数构成
4. X_i 诸变量之间无准确的线性关系，即无多重共线性	4. $\boldsymbol{X}$ 的秩 $\rho(\boldsymbol{X})=k$，其中 k 是 $\boldsymbol{X}$ 的列数并且 $k<n$
5. 为了假设检验，假定 u_i 独立、同正态分布，即 $u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$	5. 向量 $\boldsymbol{u}$ 遵从多维正态分布 $\boldsymbol{u}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I})$

纯量最小二乘剩余平方和

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \beta_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2$$

的矩阵形式为

$$\begin{aligned}\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} &= (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})\\ &= \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}.\end{aligned} \tag{1-24}$$

（注意，$\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 为一纯量）为使 $\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}=\sum \hat{u}_i^2$ 最小，将其对 $\hat{\boldsymbol{\beta}}$ 求导并令导数为零，得正规方程组

$$\begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & \cdots & \sum X_{2i}X_{ki} \\ \sum X_{3i} & \sum X_{3i}X_{2i} & \sum X_{3i}^2 & \cdots & \sum X_{3i}X_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X_{ki} & \sum X_{ki}X_{2i} & \sum X_{ki}X_{3i} & \cdots & \sum X_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

$$
=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ X_{31} & X_{32} & \cdots & X_{3n} \\ \vdots & \vdots & & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix},
$$

或简洁地写为

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

(这可直接从(1-24)式通过矩阵求导得到)两边左乘以$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$,即得$\boldsymbol{\beta}$的最小二乘估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{1-25}$$

这一解被视为LS基本理论结果的矩阵表述.这一结果不论回归方程含有多少个解释变量,也不论是否包含截距项,都一律适用.

为了对$\hat{\boldsymbol{\beta}}$进行统计推断,需要计算$\hat{\boldsymbol{\beta}}$的抽样方差协方差

$$\operatorname{var}\operatorname{cov}(\hat{\boldsymbol{\beta}}) = E\{[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]^{\mathrm{T}}\}$$

$$
=\begin{bmatrix} \operatorname{var}(\hat{\beta}_1) & \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_k) \\ \operatorname{cov}(\hat{\beta}_2,\hat{\beta}_1) & \operatorname{var}(\hat{\beta}_2) & & \operatorname{cov}(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}(\hat{\beta}_k,\hat{\beta}_1) & \operatorname{cov}(\hat{\beta}_k,\hat{\beta}_2) & \cdots & \operatorname{var}(\hat{\beta}_k) \end{bmatrix}.
$$

可以证明,在经典假设下

$$\operatorname{var}\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}, \tag{1-26}$$

其中σ^2为所有u_i的共同方差,其LS估计为

$$\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-k) = \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}/(n-k),$$

而$\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}$可由下式计算:

$$\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

(由此可得判定系数$R^2 = 1 - \dfrac{\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - n\bar{Y}^2} = \dfrac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - n\bar{Y}^2}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - n\bar{Y}^2}$)在误差遵从正态分布

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

的假设下,可以推出$\hat{\boldsymbol{\beta}}$遵从以$\boldsymbol{\beta}$为均值、$\sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$为方差的正态分布,即

$$\hat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}].$$

为了检验个别偏回归系数β_j是否显著异于β_{j0},可利用以下的t统计量:

$$t = \frac{\hat{\beta}_j - \beta_{j_0}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k},$$

对于参数的一组线性约束,则可利用下面由矩阵表示的 F 统计量

$$F=\frac{(\hat{\boldsymbol{u}}_{\mathrm{R}}^{\mathrm{T}}\hat{\boldsymbol{u}}_{\mathrm{R}}-\hat{\boldsymbol{u}}_{\mathrm{U}}^{\mathrm{T}}\hat{\boldsymbol{u}}_{\mathrm{U}})/m}{\hat{\boldsymbol{u}}_{\mathrm{U}}^{\mathrm{T}}\hat{\boldsymbol{u}}_{\mathrm{U}}/(n-k)},$$

其中下标 R 和 U 分别表示受约束和无约束的情形, m 为线性约束的个数.

1.3.2 广义最小二乘法的矩阵表述

不论回归模型中的误差项是否满足经典假设,都可以把它的方差-协方差矩阵表示为

$$\operatorname{var}\operatorname{cov}(\boldsymbol{u})=E(\boldsymbol{u}\,\boldsymbol{u}^{\mathrm{T}})=\sigma^2\boldsymbol{\Omega},$$

其中 $\boldsymbol{\Omega}$ 为 $n\times n$ 对称正定矩阵.例如,当 $\boldsymbol{u}$ 满足经典假设时, $\boldsymbol{\Omega}=\boldsymbol{I}$;当 $\boldsymbol{u}$ 属于异方差情形时,

$$\boldsymbol{\Omega}=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2\end{bmatrix}\quad(\text{不妨取 }\sigma^2=1);$$

当 $\boldsymbol{u}$ 属于一阶序列相关情形时,

$$\boldsymbol{\Omega}=\begin{bmatrix}1 & \rho & \rho^2 & \cdots & \rho^{T-1}\\ \rho & 1 & \rho & \cdots & \rho^{T-2}\\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3}\\ \vdots & \vdots & \vdots & & \vdots\\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1\end{bmatrix};$$

当 $\boldsymbol{u}$ 不满足经典假设时,将 $\boldsymbol{\Omega}$ 分解为 $\boldsymbol{\Omega}=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}$, $\boldsymbol{P}$ 为某非奇异 $n\times n$ 矩阵,然后对回归方程(1-23)作如下变量代换:

$$\boldsymbol{P}^{-1}\boldsymbol{Y}=\boldsymbol{P}^{-1}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{P}^{-1}\boldsymbol{u}.$$

其中 $\boldsymbol{P}^{-1}$ 是 $\boldsymbol{P}$ 的逆,或记为

$$\boldsymbol{Y}_*=\boldsymbol{X}_*\boldsymbol{\beta}+\boldsymbol{u}_*.\tag{1-27}$$

其中 $\boldsymbol{Y}_*=\boldsymbol{P}^{-1}\boldsymbol{Y},\boldsymbol{X}_*=\boldsymbol{P}^{-1}\boldsymbol{X},\boldsymbol{u}_*=\boldsymbol{P}^{-1}\boldsymbol{u}$.

因 $\boldsymbol{u}_*\boldsymbol{u}_*^{\mathrm{T}}=\boldsymbol{P}^{-1}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{P}^{\mathrm{T}})^{-1}=\boldsymbol{P}^{-1}\boldsymbol{\Omega}(\boldsymbol{P}^{\mathrm{T}})^{-1}=\boldsymbol{I}$,故对(1-27)式可用 LS 法.这种先将误差项变换,使它满足经典假设,再使用 LS 法的方法,就是广义最小二乘(GLS)法,且记 $\boldsymbol{\beta}$ 的广义 LS 估计为 $\boldsymbol{b}_*$,由(1-25)式和(1-26)式,有

$$\boldsymbol{b}_*=(\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{Y}_*=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{Y},$$

$$\operatorname{var}\operatorname{cov}(\boldsymbol{b}_*)=\sigma^2(\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_*)^{-1}=\sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}.$$

σ^2 的 GLS 无偏估计为

$$S^2=\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{e}}{n-k},$$

其中 $\boldsymbol{e}$ 是原回归方程的 GLS 回归残差,即

$$\boldsymbol{e}=\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}_*.$$

2 预期与分布滞后模型

预期在现代经济活动中扮演着极为重要的角色,如投资依赖于预期的利润,生产依赖于预期的销售额,长期利率依赖于预期的短期利率和预期的通货膨胀率等.

早期的预期模型,多凭直觉对预期变量的过去值进行简单的外推.例如,投资方程

$$Y_t = \alpha + \beta X_{t+1}^* + u_t. \tag{2-1}$$

其中,Y_t 为时期 t 的投资;X_{t+1}^*为预期的时期 $t+1$ 的利润;u_t 为干扰或误差.

令 X_t 为时期 t 的实际利润,由最简单的外推方法有如

$$X_{t+1}^* = X_t,$$

$$X_{t+1}^* - X_t = X_t - X_{t-1} \quad 或 \quad X_{t+1}^* = 2X_t - X_{t-1},$$

$$X_{t+1}^*/X_t = X_t/X_{t-1},$$

等等.预期(预测)值 X_{t+1}^*完全取决于模型(2-1)以外的观测值而与模型无关.至于预期效果,常用平均绝对误差 AAE 来衡量,

$$\mathrm{AAE} = \frac{1}{n}\sum_{i=1}^{n} |\,实际值_i - 预期值_i\,|.$$

进一步的外推方法将涉及预期变量的多期过去值以至全部过去值的某种加权平均,同时也就产生了估计方法的问题.先考虑用过去有限的 k 个时期取值的加权平均,作为下一时期的预期值 X_{t+1}^*的情形:

$$X_{t+1}^* = \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k}.$$

上式称为有限**分布滞后预期模型**.如何确定权数 $\beta_i(i=0,1,\cdots,k)$,特别当 k 较大时,自然是一个疑难问题.但当 $k\to\infty$ 并取 β_i 为递减的几何数列时,

$$\beta_i = \beta_0\lambda^i, \quad 0<\lambda<1, \tag{2-2}$$

却能导出一些富有实际意义的结果.

2.1 适应性预期模型

由于无穷级数

$$\sum_{i=0}^{\infty}\beta_0\lambda^i = \beta_0/(1-\lambda), \quad 0<\lambda<1.$$

取 $\beta_0 = 1-\lambda$,这样,就可把(2-2)式中的 β_i 作为 X_{t-i}项的权数,而把预期值 X_{t+1}^*表示为如下的加权平均:

$$X_{t+1}^* = \sum_{i=0}^{\infty}(1-\lambda)\lambda^i X_{t-i}. \tag{2-3}$$

将此方程滞后一期并乘以 λ,得

$$\lambda X_t^* = \lambda\sum_{i=0}^{\infty}(1-\lambda)\lambda^i X_{t-i-1} = \sum_{i=0}^{\infty}(1-\lambda)\lambda^{i+1}X_{t-i-1},$$

即

$$\lambda X_t^* = \sum_{j=1}^{\infty}(1-\lambda)\lambda^j X_{t-j}. \tag{2-4}$$

将(2-3)式减去(2-4)式,便得

$$X_{t+1}^* - \lambda X_t^* = (1-\lambda)X_t \tag{2-5}$$

或

$$\underbrace{X_{t+1}^* - X_t^*}_{\text{预期的修改}} = (1-\lambda)\underbrace{(X_t - X_t^*)}_{\text{前期的预期误差}}. \tag{2-6}$$

(2-6)式表明,要根据前期预期的误差来修改下一期的预期.顾名思义,所以把这种预期方式名为**适应性(或自适应)预期模型**(adaptive expectations model),模型中 $0<\lambda<1$ 表明,预期修改的幅度($X_{t+1}^* - X_t^*$)比预期误差($X_t - X_t^*$)要小些.

由于模型(2-3)中的全部系数总和为1,如果 X_t 有上升的长期趋势,预期值 X_{t+1}^* 还应乘以$(1+g)$,其中 g 代表 X_t 的平均增长率.否则,预期模型便会系统地低估 X_t 的真实值.

把(2-5)式同(2-1)式联系起来即可消去(2-1)式中不能直接观测的 X_{t+1}^* 项.为此,将(2-1)滞后一期,乘以 λ,然后从(2-1)式减去,便得

$$\begin{aligned} Y_t - \lambda Y_{t-1} &= \alpha(1-\lambda) + \beta(X_{t+1}^* - \lambda X_{t-1}^*) + u_t - \lambda u_{t-1} \\ &= \alpha(1-\lambda) + \beta(1-\lambda)X_t + u_t - \lambda u_{t-1} \end{aligned}$$

或

$$Y_t = \alpha' + \lambda Y_{t-1} + \beta' X_t + v_t. \tag{2-7}$$

其中 $\alpha' = \alpha(1-\lambda)$, $\beta' = \beta(1-\lambda)$, $v_t = u_t - \lambda u_{t-1}$.

(2-7)式就是从适应性预期假说(2-6)式导出的一个回归模型.它含有一项滞后一期的应变量 Y_{t-1} 作为解释变量.凡是含有滞后一期或多期滞后应变量的回归方程都称为**自回归模型**(autoregressive model),也称**动态**(dynamic)**模型**,它描述了应变量是怎样同它自己过去发生关系的**时间走道**(time path).(2-7)式是一个动态模型,其中 X_t 的系数 β' 仅代表一种短期的投资行为,而 Y_{t-1} 的系数 λ 则反映了一种长期的投资效应.参看下例.

例1 短期与长期总消费函数

设消费 C_t 和永久收入 X^* 有如下的线性关系:

$$C_t = \alpha + \beta X_{t+1}^* + u_t. \tag{2-8}$$

按照适应性预期模型(2-5)式,可导出一个形如(2-7)式的方程:

$$C_t = \alpha' + \beta' X_t + \lambda C_{t-1} + v_t,\ v_t = u_t - \lambda u_{t-1}. \tag{2-9}$$

假定通过适当的估计方法得到①

$$\hat{C}_t = 2.361 + 0.2959\ X_t + 0.6755\ C_{t-1}.$$

标准误 (1.229) (0.0582)

① 利用 M.C. Lovell, Macroeconomics: Measurement, Theory and Policy, 1975, 第148页的数据.

这一数值结果表示:**边际消费倾向**(MPC)为 0.2959,即当年可支配收入 X_t 每增加 1 元,消费便即时增加约 0.3 元. 如果收入的增加是持久的. 则消费最终增加 $\beta = \beta'/(1-\lambda) = 0.2959/(1-0.6755) \approx 0.91$ 元. 这点也可直接从(2-9)式看出. 当 t 很大时,也就是从长期看消费达到均衡时,C_{t-1} 可视同 C_t,即可令 $C_{t-1} = C_t$,于是从(2-9)式解出

$$C_t = \frac{\alpha}{1-\lambda} + \frac{\beta'}{1-\lambda}X_t + \frac{v_t}{1-\lambda}.$$

其中系数 $\beta'/(1-\lambda) = \beta$ 就成为相对于(2-8)式中的 X_t^* 的长期 MPC,而 β' 仅是相对于(2-9)式中的 X_t 的短期(即期)MPC.

2.1.1 部分调节模型

一个酷似模型(2-7)的形式,还可从另一种经济行为的描述得到. 人们考虑如何使自己的行为达到最优,往往要经过一个调节过程. 例如,石油价暴涨,如何将消费量调节到最优水平? 从短期看可减少开车率,降低取暖标准等等;从长期看可提高石油燃烧率,开发新能源等等. 故消费量一时只能做到部分调节. 设相应于油价 X_t 的最优消费 Y_t^* 为

$$Y_t^* = \alpha + \beta X_t + u_t, \tag{2-10}$$

部分调节(partial adjustment)过程为

$$\underbrace{Y_t - Y_{t-1}}_{\text{实际调节}} = \delta\,(\underbrace{Y_t^* - Y_{t-1}}_{\text{最优调节}}), \quad 0<\delta\leqslant 1. \tag{2-11}$$

上式表明,实际调节仅为最优调节的一个分数,其大小 δ 称为调节系数. 将(2-10)式代入(2-11)式可得

$$Y_t = \alpha\delta + (1-\delta)Y_{t-1} + \beta\delta X_t + \delta u_t, \tag{2-12}$$

这就是根据部分调节假说(2-11)式导出的回归模型. 同模型(2-7)相比,(2-12)式的误差项较为简单;(2-7)式的误差项明显是自相关的,而(2-12)式则未必如此. 因此还有必要对(2-7)式考虑不同的估计方法.

类似于适应性预期模型之用于长、短期分析,对部分调节模型来说,(2-10)式代表一种长期消费行为,其长期 MPC 为 β,而(2-12)式则描述了一种短期消费行为,其短期 MPC 为 $\beta\delta$,后者要除以 δ 才给出长期 MPC,δ 是调节系数,它表明消费者在时间走道上迈向最优的长期消费水平的每一时期的进程.

2.1.2 工具变量法

在回归模型中,只有滞后应变量(作为解释变量)而无时序相关干扰,或只有时序相关干扰而无滞后应变量时,普通最小二乘(OLS)估计量是相合或一致性的. 但兼有二者时(如在适应性预期模型中那样),OLS 估计量是不相合的. **工具变量**(instrumental variables,简记 IV)法是解决估计的**相合性问题**的一个主要方法.

用矩阵符号表示,OLS 估计 $\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 可看作是用 $\boldsymbol{X}^{\mathrm{T}}$ 左乘回归方程 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{u}$,使

$$Y = X\beta + u \Rightarrow X^{T}Y = X^{T}X\beta + X^{T}u,$$

并在 $X^{T}X$ 非奇异假设下令 $X^{T}u=0$ 得到的结果,故可把 X 看作一种求解的"工具".当回归模型,比方说,适应性预期模型

$$Y_t = \alpha + \beta X_t + \lambda Y_{t-1} + v_t, \quad v_t = u_t - \lambda u_{t-1}$$

中的 $X_t = [X_t\ Y_{t-1}]$不适合作为求解参数的"工具"时,是否可以另找其他变量代替 X_t? 鉴于在 $X=[X_1, X_2, \cdots, X_n]^{T}$ 适合作为最小二乘法求解的"工具"的情况下,有

(1) $E(X^{T}u)=0$ 且 $X^{T}u/n \xrightarrow{p} 0$;

(2) $\mathrm{var}(b)=\sigma^2(X^{T}X)^{-1}$;

(其中"$\xrightarrow{p}$"表示"概率意义下趋于")因此,要寻找工具变量(IV) $Z=[Z_1\ Z_2]$,使得

(1′) $\mathrm{plim}\left(\frac{1}{n}Z_1^{T}v\right)=0$, $\mathrm{plim}\left(\frac{1}{n}Z_2^{T}v\right)=0$ (鉴于(1));

(2′) $\mathrm{plim}\left(\frac{1}{n}Z_1^{T}X\right)$ 及 $\mathrm{plim}\left(\frac{1}{n}Z_2^{T}X\right)$ 均为异于零的有限值,且最好是 Z 与 X 高度相关(鉴于(2)).

作为一种建议,可取 $Z_{1t}=X_t, Z_{2t}=X_{t-1}$.对适应性预期模型来说,就是用 X_{t-1} 代替 Y_{t-1}作为回归估计的工具变量.

记工具变量法的估计量为 b_{IV},将有

$$b_{IV}=(Z^{T}X)^{-1}Z^{T}Y, \quad \mathrm{plim}\, b_{IV}=\beta, \tag{2-13}$$

后一等式表明 b_{IV}是相合或一致性估计.

2.1.3 含滞后应变量的自相关检验

在部分调节模型(2-12)式中,如果 u_t 无时序相关,$v_t=\delta u_t$ 也就无时序相关;而在适应性预期模型中,即使 u_t 无时序相关,$v_t=u_t-\lambda u_{t-1}$仍有时序相关.由于自回归模型含有滞后应变量,若用 D-Wd 统计量检验一阶时序相关,则 d 值有偏近于 2 的倾向,从而导致过多地接受无时序相关的零假设 H_0,也就是增大了第Ⅱ类错误的概率.为了寻求适当的检验统计量,德宾(J. Durbin)得到了一个大样本结果:在自回归模型中,可用 h 统计量代替 d 统计量,h 和 d 的关系如下:

$$h=\hat{\rho}\sqrt{\frac{n}{1-n(\mathrm{var}(\hat{\alpha}_2))}}, \quad \hat{\rho}\approx 1-\frac{1}{2}d$$

或

$$h\approx\left(1-\frac{1}{2}d\right)\sqrt{\frac{n}{1-n(\mathrm{var}(\hat{\alpha}_2))}}.$$

其中 n 为样本大小,$\hat{\alpha}_2$ 为回归模型中 Y_{t-1}的系数 α_2 的最小二乘估计.德宾证明,当 n 很大时,h 渐近地遵从标准正态分布.

$$h \overset{\text{asy}}{\sim} N(0,1).$$

不论回归模型中含有多少个自变量和多少期滞后应变量，这个渐近的正态分布都成立．这样就可以利用正态概率，比如用

$$Pr(-1.96 \leqslant h \leqslant 1.96) = 0.95$$

作出判断：当$|h| > 1.96$时，以显著水平$(1-0.95) = 0.05$拒绝无一阶自相关的零假设H_0；否则；当$|h| \leqslant 1.96$时，接受H_0.

当$n\mathrm{var}(\widehat{\alpha_2}) > 1$时，$h$检验便不适用．这时，德宾建议先对自回归方程求LS误差估计值$\widehat{u_t}$，再做$\widehat{u_t}$对$\widehat{u_{t-1}}$，Y_{t-1}和X_t的回归，然后根据$\widehat{u_{t-1}}$的系数异于零的显著性，以判断H_0：自相关系数$\rho=0$可否接受．

2.1.4　多项式滞后

以上主要讨论无穷滞后分布，并且滞后项的权数分布是按几何级数递减的．下面讨论一种较常应用的有限滞后分布——**多项式滞后**(polynomial lag)，它的滞后项的权数是按一个设定的多项式拟合的．不论原来回归模型含有多少个滞后项，都能把它减少到同所设多项式的次数一样多(甚至更少)的项，从而避免滞后项较多时回归估计中扰人的多重共线性问题．

例2　对含s个滞后项的自回归模型(不妨看作对投资方程(2-1)的另一种选择)

$$Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-s} + u_t \tag{2-14}$$

中的β_i拟合一个r次多项式

$$\beta_i = f(i) = a_0 + a_1 i + a_2 i^2 + \cdots + a_r i^r, \tag{2-15}$$

比方说，取$r=2$，$s=7$，于是得

$$\beta_0 = a_0, \beta_1 = a_0 + a_1 + a_2, \cdots, \beta_7 = a_0 + 7a_1 + 49a_2,$$

将其代入待估方程(2-14)得

$$\begin{aligned} Y_t = {} & a_0(X_t + X_{t-1} + X_{t-2} + \cdots + X_{t-7}) + \\ & a_1(X_{t-1} + 2X_{t-2} + 3X_{t-3} + \cdots + 7X_{t-7}) + \\ & a_2(X_{t-1} + 4X_{t-2} + 9X_{t-3} + \cdots + 49X_{t-7}) + u_t. \end{aligned} \tag{2-16}$$

这样，对8个系数$\beta_i(i=0,1,2,\cdots,7)$的估算就减为对3个系数$a_i(i=0,1,2)$的估算．但关键是要在应用中把最可能长的滞后期$s$和多项式的次数$r$定得适当．

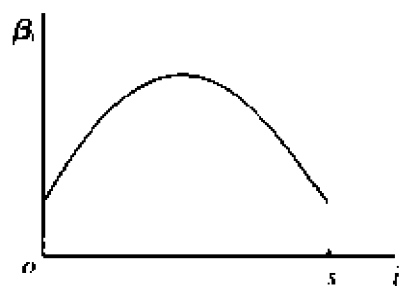


图2-1　多项式分布滞后

从(2-15)式看出，β_i随r次多项式的值而变化，多项式滞后是否具有合理性就在于它所描绘的几何形状是否代表各滞后项应有的权数．不过，一个同次(比方说2次)多项式可以有许多不同的形状，可以通过对β_i的线性约束(线性约束便于统计检验)，适当地限制它的几何图形，使之更为合理．最常用的约束是端点约束．例如对

$$\beta_i = f(i) = a_0 + a_1 i + a_2 i^2,$$

限定$\beta_{-1}=0$及$\beta_{s+1}=0$(参看图2-1)就得到

$$f(-1) = a_0 - a_1 + a_2 = 0 \text{ 及 } f(s+1) = a_0 + a_1(s+1) + a_2(s+1)^2 = 0.$$

由此给出如下约束关系：

$$a_0 = -a_2(s+1) \text{及} a_1 = -a_2 s,$$

从而把(2-16)式简化为

$$Y_t = a_2 Z_t + u_t.$$

其中

$$Z_t = \sum_{i=0}^{s} (i^2 - si - s - 1) X_{t-i} \quad (s=7).$$

以上简例说明了多项式滞后在应用中还有许多灵活性.

多项式滞后又称阿尔蒙(Almon)滞后.

2.2 合理预期

适应性预期用于研究某些带有长期趋势的变量如通货膨胀率等，是不能令人满意的.如果价格缓慢地或升或降，一味按上期价格水平进行部分修改的适应性预期，则尚可理解.但当价格出现一次性跳动或兼有较强的长期增长或衰退趋势时，缓慢改动的适应性预期其误差将越变越大.合理的预期应参考过去更多的价格甚至全部价格历史作为依据，而不局限于仅仅过去一期的价格数据.不仅如此，还应参考直至现在为止的全部过去的有关信息作为预期的条件.所谓有关信息，不仅包括价格信息，还包括其他有关变量，至少包括全部价格历史.

记直至前一时期末的信息集为 I_{t-1}.简言之，**合理预期**(rational expectation)就是以 I_{t-1} 为条件的条件预期或条件数学期望 $E(\cdot \mid I_{t-1})$，其中"·"代表所预期的变量.合理的预期应有如下性质：

(1) 预期(值)应是无偏的.比如，记时期 t 的价格 Y_t 的预期(值)为 Y_t^*，预期误差为 ϵ_t，预期的无偏性是指

$$Y_t = Y_t^* + \epsilon_t, \quad E(\epsilon_t) = 0. \tag{2-17}$$

如果把 Y_t 看作随机的，则对 Y_t 的"主观"预期 Y_t^* 就等于它的"客观"均值：

(2) $$Y_t^* = E(Y_t \mid I_{t-1}), E(\epsilon_t \mid I_{t-1}) = 0. \tag{2-18}$$

这也可看作对(2-17)式两边分别求客观数学期望的结果.(2-18)式现已成为所有有关合理预期的计量经济研究的基础.

合理预期的无偏性要求(unbiasedness requirement)的一个重要含义是，对所有未来时期的预期来说，预期误差的期望值均为零.比如，对提前一期和提前二期的合理预期应有

$$\begin{cases} Y_{t+1} = Y_{t+1}^* + \epsilon_{1,t+1} \\ \qquad = E(Y_{t+1} \mid I_t) + \epsilon_{1,t+1}, E(\epsilon_{1,t+1} \mid I_t) = 0, \\ Y_{t+1} = E(Y_{t+1} \mid I_{t-1}) + \epsilon_{2,t+1}, E(\epsilon_{2,t+1} \mid I_{t-1}) = 0. \end{cases} \tag{2-19}$$

其中，$\epsilon_{i,t+i}$ 表示提前 i 期的预期误差，并假定 ϵ_1 与 ϵ_2 相互独立，ϵ_1 与 I_t 以及 ϵ_2 与 I_{t-1} 均不相关(这个"不相关"的含义将在下面说明).将(2-19)式中的两等式相减再求条件期望，得

$$E\{[E(Y_{t+1}|I_t)-E(Y_{t+1}|I_{t-1})]|I_{t-1}\}=0.$$

因 ϵ_1 及 ϵ_2 的条件期望为零，由此导出有重要意义的所谓**叠期望律**(law of iterated expectations)：

(3)　$E[E(Y_{t+1}|I_t)|I_{t-1}]=E(Y_{t+1}|I_{t-1})$.　　(2-20)

叠期望律表明，在合理预期假设下，不仅所有未来时期的预期是无偏的，而且一切预期(值)的预期也是无偏的. 在金融学史上有一个著名的叠期望律的应用成果，这就是萨缪尔逊(P. Samuelson)1965年通过叠期望律证明了"适当地预测的证券价格是随机地波动的"这一命题.

叠期望律的一个更一般的表述形式为

$$E(X|I_t)=E[E(X|J_t)|I_t],\ I_t\subset J_t.$$

这里，$I_t\subset J_t$ 表示信息集 J_t 蕴含着 I_t(例如上面的 $I_{t-1}\subset I_t$). 若把对 X 的合理预期看作最优预测，则叠期望律又可解释为

对 X 的最好的预测总等于对 X 更好的预测(值)的最好预测. (因为 J_t 包含 I_t，所以 $E(X|J_t)$ 是比 $E(X|I_t)$ 更好的预测.)

的确，用 I_{t-1} 作条件预测的目的，正是为了作出"最优"预测. 所以在预测中要充分利用 I_{t-1}，这就意味着预期误差 ϵ_t 应与 I_{t-1} 无关；否则还可以利用 ϵ_t 和 I_{t-1} 之间的相关性以改进预测.

(4) 假定 ϵ_t 与上面(2-18)式中的 I_{t-1} 无关，那么(2-17)式中的 Y_t^* 与 ϵ_t 无关. 对(2-17)式两边求方差，得

$$\mathrm{var}(Y_t)=\mathrm{var}(Y_t^*)+\mathrm{var}(\epsilon_t).$$

因方差 $\mathrm{var}(\epsilon_t)$ 非负，故

$$\mathrm{var}(Y_t)\geqslant\mathrm{var}(Y_t^*).\qquad(2\text{-}21)$$

预期(值) Y_t^* 有较小的方差，合理预期的"最优"含义就在此.

以上无偏性和较小方差性是合理预期假设(rational expectation hypothesis)的两个基本性质. 至于主观预期等于客观(总体)平均以及叠期望律，则可视为无偏性的深一层推论. 仅要求无偏性的合理预期定义称为**弱式(weak form)合理预期**，兼要求较小方差的合理预期称为**强式(strong form)合理预期**. 弱式和强式合理预期假设都在金融学或经济学中有过重大的应用成果.

2.2.1　应用举例：预期无偏性与政策无效性推理

现代新古典经济学与合理预期的结合，产生如下的总供给与总需求模型：

总供给 SS：

$$Y_t=Y_{t-1}+\beta(P_t-P_t^*)+\epsilon_t.\qquad(2\text{-}22)$$

总需求 DD：

$$P_t=-\alpha(Y_t-Y_{t-1})+M_t+u_t.\qquad(2\text{-}23)$$

上两式中，P_t^*(或记 $P_{t|t-1}^*$)表示在时期 $t-1$ 末对价格 P_t 的预期，无碍于一般性，假定 P_{t-1} 已单位化为 $P_{t-1}=1$；Y_t 为国民产出或收入；M_t 为货币量(代表政策变量)，并取 $M_{t-1}=1$；α 和 β 为参数；ϵ_t 和 u_t 为误差.

把 P_t^* 和 M_t 看作预先给定,即可将 P_t 消去而解出 Y_t(称 Y_t 的诱导或约简式,参看下一章 3.3 节),

$$Y_t = Y_{t-1} + \frac{\beta}{1+\alpha\beta}(M_t - P_t^*) + \frac{1}{1+\alpha\beta}(\beta u_t + \epsilon_t). \tag{2-24}$$

怎样决定 P_t^*? 如按照适应性预期,则即使看到 M_t 这个政策变量在变,也仅仅是根据已发生的预期误差来改变 P_t^*,可以说这是一种面向过去的方法.这种方法仅当价格一直在围绕某均衡平均水平而随机地波动时才有意义.而如果按照合理预期,则应看到 M_t 的未来变化而加以预测,所以它是一种面向未来的方法.这就需要根据理论模型来决定 P_t^*.

为此,将(2-23)式中的时期提前一期考虑.因为在时期 $t-1$ 尚不知道时期 t 的实际值,故用预期值代替实际值,得

$$P_t^* = -\alpha[Y_t^* - Y_{t-1}] + M_t^*. \tag{2-25}$$

其中 u_t 消失是因为它的合理预期为零.

再对(2-22)式作合理预期.由于$(P_t - P_t^*)$和 ϵ_t 的合理预期均为零,故对 Y_t 的合理预期为

$$Y_t^* = Y_{t-1}. \tag{2-26}$$

将(2-26)式代入(2-25)式便得

$$P_t^* = -\alpha(Y_{t-1} - Y_{t-1}) + M_t^* = M_t^*. \tag{2-27}$$

这是合理或内生(即由模型内导出的)预期的一个有深刻意义的结果.因为,将(2-27)式代入(2-24)式中的 P_t^* 可得

$$Y_t = Y_{t-1} + \frac{\beta}{1+\alpha\beta}(M_t - M_t^*) + \frac{1}{1+\alpha\beta}(\beta u_t + \epsilon_t). \tag{2-28}$$

(2-24)式和(2-28)式同是 Y_t 的诱导式,但其经济含义迥异.在(2-24)式中,产出的变化由 M_t 和 P_t^* 的差异决定.按照适应性预期,P_t^* 仅由过去的价格决定,而过去的价格怎样地同现在的货币政策发生联系,就难于解释.但在(2-28)式中,产出的变化完全由 M_t 和 M_t^* 的差异决定.如果预期是合理的,M_t^* 应接近于并且平均而言等于 M_t,产出的变化就会很小.

因此得出结论:不论政策变量 M_t 怎样变,如果它的变化完全为人们所意料,这种政策对产出的影响将是零,即 $Y_t = Y_{t-1}$,这就是现代合理期望学派所谈论的著名的"政策无效性"定理的基本依据.只有非预期的政策(或政策的非预期部分)才能对产出的改变起作用.

上述"政策无效性"定理所依据的仅是合理预期的无偏性,即弱式合理预期假设.即使考虑强式合理预期,"政策无效性"结论在平均意义上仍然成立,只不过误差项$(\beta u_t + \epsilon_t)/(1+\alpha\beta)$的方差从(2-24)式到(2-28)式会有所改变而已.①

① 参见:Branson W H. Macroeconomic Theory and Policy. 3rd. ed. New York: Harper & Row, 1992.

2.2.2 合理预期假设的检验

1. 弱式(无偏性)检验

通常能从市场调查资料中直接询问或收集到 Y_t 的预期值 Y_t^*. 预期的无偏性要求预期误差 $\varepsilon_t = Y_t - Y_t^*$ 与信息集 I_{t-1}无关. 因 Y_{t-1}必定属于 I_{t-1},故可通过回归方程

$$Y_t - Y_t^* = \alpha_0 + \alpha_1 Y_{t-1} + u_t, \tag{2-29}$$

检验假设 $H_0: \alpha_0 = 0, \alpha_1 = 0$. 无偏性意味着 $\alpha_1 = 0$. 拒绝 H_0 即表示否定预期(值)的无偏性.

还可估计另一个回归方程

$$Y_t - Y_t^* = \alpha_0 + \alpha_1(Y_{t-1} - Y_{t-1}^*) + u_t. \tag{2-30}$$

看预期误差 u_t 是否有非零均值和是否与过去的预期误差无自相关. 无偏性要求自相关系数 $\alpha_1 = 0$. 如果通过假设检验认为 $\alpha_1 \neq 0$,则表明在预期中 I_{t-1}中的信息未得到充分利用.

然而,在做回归(2-29)式或(2-30)式之前,又不妨先尝试一个更简单的回归方程

$$Y_t = \beta_0 + \beta_1 Y_t^* + u_t, \tag{2-31}$$

以检验 $H_0: \beta_0 = 0, \beta_1 = 1$,看预期误差是否平均为零(参看 5.3.1 节).

2. 强式(较小方差性)检验

当 $Y_t = Y_t^* + \varepsilon_t$ 且 ε_t 与 Y_t^* 无关时,将有

$$\mathrm{var}(Y_t) = \mathrm{var}(Y_t^*) + \mathrm{var}(\varepsilon_t) \geqslant \mathrm{var}(Y_t^*).$$

强式检验以不等式 $\mathrm{var}(Y_t) \geqslant \mathrm{var}(Y_t^*)$是否成立作为依据. 这个检验又可分别为对 $\mathrm{var}(Y_t) = \mathrm{var}(Y_t^*)$的正交性($Y_t^* \perp \varepsilon_t$)检验和对 $\mathrm{var}(Y_t) > \mathrm{var}(Y_t^*)$的边界检验.①

在弱式检验中不免有一个疑问,即在信息集 I_{t-1}中一般都有多个变量,比方说,变量 Z_1 和 Z_2 都属于 I_{t-1},即使($Y_t - Y_t^*$)对 Z_1 的回归不显著,但如果 Z_2 与 Z_1 无关,($Y_t - Y_t^*$)对 Z_2 的回归仍可能是显著的. 那么,仅凭 Z_1(比如说 Y_{t-1})的信息就可作出预期是否合理的判断? 这似有问题. 这个问题一方面确实存在,但另一方面,不难想象,I_{t-1}中的变量虽多,但大多数情形下这些变量都是相关的,能抓住主要变量问题就不大.

洛弗尔(M.C.Lovell)在 1986 年利用市场调查资料,对销售、库存、价格、工资等的预测数据做过许多弱式和强式的合理预期假设检验,发现大多数情形的预测值都不符合合理预期假设.

3 联立方程模型

在单方程回归模型中,因果关系被认为是单向的,即从一个或多个解释变量 x

① 参阅 Jourl. of Economic Literature, Dec. 1989, vol. 27, 第 1595 ~ 1603 页; American Economic Review, Mar. 1986, vol. 76, 第 110 ~ 124 页.

到一个被解释的应变量 y.但更多的情形是,同时有多个方程,y 一方面决定于 x,另一方面某些 y 又决定于另一些 y,简言之,诸 y 之间的因果关系是双向或相互的.这尤其因为经济数据是一种非实验数据,一般不能人为地固定一些变量而去观测另一些变量.若把应变量和解释变量截然分开,则是不切实际的.例如,工资与价格是互相影响的,市场上的价格与成交量也是同时决定的,等等.因此,合理的做法将是用一组联立方程(simultaneous equations system)去描述变量之间的相互依从关系,使其中的一组变量(内生变量)在给定其余变量(外生变量)的情况下同时被决定.

3.1 结构方程

在联立方程中,每个方程描述一种经济结构,故**联立方程模型**又称为结构方程组(system of structural equations)模型,或简称结构模型.其中每个方程均称为**结构方程**.它可以是技术性的,如生产方程;也可以是行为性的,如消费方程;还可以是定义性的,如会计上的恒等式.下面通过两个例子来阐明这种结构模型的性质,其中一个涉及横截面数据,另一个则涉及时间序列数据.

例 1 在市场经济中,女职工每周的劳动时数和领取的工资都是有一定变化的.从一个雇主的观点看,考虑到职工的责任感和连续工作的重要性,将劳动时数适当延长,每小时劳动的价值会大些.由此可导出他所需求的工作时数与他实际支付的工资之间的一种劳动需求关系(3-1)式.而从一个职工的观点看,她愿意提供的工作小时数和工资报酬、家庭其他收入以及抚养的儿童人数有关.由此可导出她的劳动供给关系(3-2)式.把这两方面的行为合并起来考虑,即得到一个结构模型:

需求: $$\text{工时}_i = \beta_0 + \beta_1\,\text{工资}_i + u_i, \tag{3-1}$$

供给: $$\text{工时}_i = \gamma_0 + \gamma_1\,\text{工资}_i + \gamma_2\,\text{其他收入} + \gamma_3\,\text{儿童数} + v_i. \tag{3-2}$$

在此模型中,工时既是需求量又是供给量,作为实际工作时数,这两个量是在**观测上等效**(observationally equivalent)的,但在概念上须区别为工时^d(需求量)和工时^s(供给量),只不过工时^d = 工时^s 而已.为简单起见,(3-1)式和(3-2)式省掉了上标 d 和 s,也就省掉了一个等式:工时^d = 工时^s.

联立方程模型的分析,需要对变量作新的分类.因为工时和工资这两个变量由系统内部(指方程组本身)决定,故称**内生**(endogenous)**变量**,而其他收入和儿童数决定于系统之外,故称**外生**(exogenous)**变量**.为了突出需求和供给两种不同行为,比较自然的表达式是把工时这个变量都放在方程的左端.事实上,工时和工资作为内生变量有完全等同的地位的.

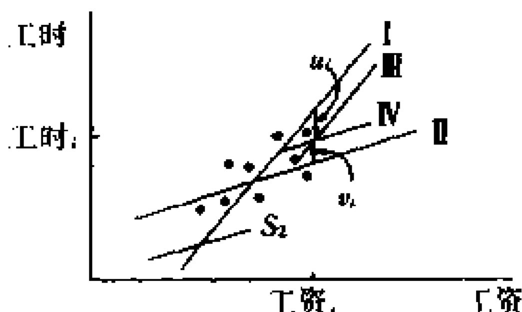


图 3-1 工时工资同时被决定

为了说明内生变量工时和工资是怎样同时被决定的,不妨考虑其他收入和儿童数都取固定值的一组关于工时和工资的观测值(由图 3-1 中的散点表示).这时(3-2)式中的 γ_2(其他收入)和 γ_3(儿童数)都是常数.

如果取干扰(误差)项 $u_i=0, v_i=0$,就可画出代表(3-1)式和(3-2)式的系统部分的两条直线,如图3-1中线段Ⅰ和Ⅱ.两线的交点同时决定了工资和工时.从雇主的观点看,工资越高,表明对工作小时数的需求越大;从工人的观点看,工资越高,表明对工作小时数的供给越多,故两个斜率 β_1 和 γ_1 都取正值.

由于干扰的随机性,每次观测的结果都不一样.如果第 i 次观测的 u_i 为负而 v_i 为正,如图3-1所示,那么,对这个观测点来说,工时-工资的需求关系式(其截距和斜率就分别为 β_0+u_i 和 β_1,且将位于系统的需求关系式Ⅰ的下方,其距离相当于负数 u_i.同理,对这个观测点来说,供给关系式位于它的系统关系式Ⅱ之上.这两个关系式由图中通过观测点 i 的两条线段Ⅲ和Ⅳ表示.在两线交点处,两个方程都被满足,即工时和工资再次同时被决定.对于其他的观测点,均可作类似的解释.

当其他收入和抚养儿童数取另外两个给定值时,就会得到另一条供给曲线如图3-1中的 S_2.比如说,S_2 代表有更多其他收入和抚养更多儿童的女职工工时供给线,则不论现行工资如何,她愿意工作的小时数都要少些.

例2 根据凯恩斯理论,一个简单的国民收入宏观(总量)模型如下:

假定是一个无外贸的封闭经济系统,则有恒等关系式

$$Y=C+i+g. \tag{3-3}$$

其中,Y 为国民收入(GNP),C 为消费支出,i 为投资额,g 为政府支出.为了解释其中的构成部分,通过经济行为的分析,认为消费支出是税后收入和利息率的函数,而投资是GNP的变化和利息率的函数.假定函数是线性的,有

$$C_t=\alpha_0+\alpha_1(1-\tau)Y_t+\alpha_2\gamma_t+u_t, \tag{3-4}$$

$$i_t=\beta_0+\beta_1\Delta Y_{t-1}+\beta_2\gamma_{t-1}+v_t, \quad \Delta Y_{t-1}=Y_{t-1}-Y_{t-2}. \tag{3-5}$$

其中 τ 为税率;γ 为利息率;u 和 v 代表误差;下标表示时期;参数 α_1 表示边际消费倾向;β_1 代表投资加速作用,并且预料

$$0<\alpha_1<1, \alpha_2<0, \beta_1>0, \beta_2<0.$$

为了注明(3-3)式相对于(3-4)式和(3-5)式的时间,可将(3-3)式写为

$$Y_t=C_t+i_t+g_t, \tag{3-6}$$

至于利息率 γ 和政府支出 g 是怎样决定的,就不再追问,而认为是外生的.如果还有哪个外生变量被认为不是外生而需要解释,则要为它再增设一个关系式,从而把它内生化.现在既然把 γ 和 g 看作外生,方程(3-4)、(3-5)和(3-6)就构成一个完整的联立方程模型.其中结构方程的个数等于内生变量的个数(本例中都是3).一旦给定了外生变量的值,即可通过解联立方程组(3个变元的3个线性方程组),定出全部三个内生变量的值.

3.1.1 普通最小二乘法与联立性偏误

最小二乘回归的优良性质BLUE有一个重要的前提假设,即解释变量与误差无关.在联立方程模型中,作为解释变量的内生变量一般都与误差项有相关关系.例如,在例2的结构方程(3-4)中,作为解释变量的 Y_t,由于(3-6)式的关系,是一个内生变量,便和误差 u_t 有关.若直接用OLS法估计结构系数 β_1,则这种估计不但不

是 BLUE，而且是非一致性(inconsistent)估计.这就是说，不管样本多大，都消除不了估计的偏误.为了确切地说明这点，作为示例，不妨把例 2 中的模型再进一步简化.

假定投资 i 也是外生的.这时(3-5)式可以略去，i 和 g 就没有什么区别了，而可将(3-6)式写成

$$Y_t = C_t + Z_t \quad (Z_t = i_t + g_t). \tag{3-7}$$

再记(3-4)式中的 $\alpha_0 = \alpha, \alpha_1(1-\tau) = \beta, \alpha_2 = 0$，则(3-4)式可简记为

$$C_t = \alpha + \beta Y_t + u_t, \tag{3-8}$$

从而(3-7)式和(3-8)式构成了一个更为简单的联立方程模型.收入 Y_t 在(3-8)式中是内生解释变量，但在(3-7)式中却是内生被解释变量.只要把(3-8)式中的 C_t 代入(3-7)式，就能看出 Y_t 与 u_t 相关.具体步骤是

$$Y_t = C_t + Z_t = \alpha + \beta Y_t + u_t + Z_t,$$

由此得

$$Y_t = \frac{\alpha}{1-\beta} + \frac{Z_t}{1-\beta} + \frac{u_t}{1-\beta},$$

$$E(Y_t) = \frac{\alpha}{1-\beta} + \frac{Z_t}{1-\beta},$$

$$Y_t - E(Y_t) = \frac{u_t}{1-\beta},$$

$$\operatorname{cov}(Y_t, u_t) = E\{[Y_t - E(Y_t)]u_t\} = \frac{E(u_t^2)}{1-\beta} = \frac{\sigma_u^2}{1-\beta} > 0.$$

可见 Y_t 和 u_t 有明确的相关关系.

记 β 的 LS 估计为 $\widehat{\beta}$，则

$$\widehat{\beta} = \frac{\sum c_t y_t}{\sum y_t^2} = \frac{\sum C_t y_t}{\sum y_t^2} \qquad \left(c_t = C_t - C, C = \frac{\sum_{t=1}^{n} C_t}{n}\right),$$

将(3-8)式代入上式得

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum(\alpha + \beta Y_t + u_t)y_t}{\sum y_t^2} \\
&= \frac{\alpha\sum y_t + \beta\sum Y_t y_t + \sum u_t y_t}{\sum y_t^2} = \frac{0 + \beta\sum y_t^2 + \sum u_t y_t}{\sum y_t^2} \\
&= \beta + \frac{\sum u_t y_t/n}{\sum y_t^2/n} \xrightarrow{p} \beta + \frac{\sigma_u^2/(1-\beta)}{\sigma_Y^2} > \beta.
\end{aligned}$$

(因 $0 < \beta < 1$).其中"$\xrightarrow{p}$"表示"概率上趋于".可见 $\widehat{\beta}$ 过大估计了 β.无论 n 多大，这个过大的偏误都不会消失，这种偏误是由方程的联立性引起的，故称之为(LS)**联立性偏误**(simultaneity bias).

3.1.2 诱导方程与间接最小二乘法

为了得到 β 的一致性估计，可考虑间接最小二乘(indirect least squares，简记 ILS)法：先从模型中把待估的内生变量解出来，得到所谓诱导(型)方程(reduced (form) equation)，再用 LS 法估计诱导方程的系数，即所谓倍(乘)数(multiplier)，然后通过倍数与结构系数的代数关系式，求出结构系数．例如，由(3-8)式和(3-7)式构成的结构模型，从中消去 Y_t，即解出 C_t 的诱导方程为

$$C_t = \Pi_0 + \Pi_1 Z_t + w_t. \tag{3-9}$$

其中

$$\Pi_0 = \frac{\alpha}{1-\beta}, \quad \Pi_1 = \frac{1}{1-\beta}, \quad w_t = \frac{u_t}{1-\beta}. \tag{3-10}$$

诱导方程(3-9)的系数 Π_1 惯称倍数．由于(3-9)式中的解释变量 Z_t 是外生变量，与误差项 w_t 无关，故可用 OLS 法估计参数 Π_0 和 Π_1．再由于 LS 估计量 $\widehat{\Pi_0}$ 和 $\widehat{\Pi_1}$ 的一致性，通过代数关系式(3-10)，将估计值 $\widehat{\Pi_0}$ 和 $\widehat{\Pi_1}$ 代入 Π_0 和 Π_1，便可求出 α 和 β 的一致估计．

但是，并非任何模型的结构方程都能通过 ILS 法或其他方法求得结构系数的一致性估计．不妨看看上述例 1 中的供给方程(3-2)．为说明简单起见，把供给方程(3-2)中的抚养儿童项划掉，来考虑一个较简单的供求模型，即

需求：
$$\text{工时}_i = \beta_0 + \beta_1 \text{工资}_i + u_i, \tag{3-11}$$

供给：
$$\text{工时}_i = \gamma_0 + \gamma_1 \text{工资}_i + \gamma_2 \text{其他收入}_i + v_i. \tag{3-12}$$

注意"其他收入"是给定的外生变量．由解二元联立方程的标准代数方法能得到

$$\text{工资}_t = \Pi_{10} + \Pi_{11}\text{其他收入}_t + \epsilon_{1t}, \tag{3-13}$$

$$\text{工时}_t = \Pi_{20} + \Pi_{21}\text{其他收入}_t + \epsilon_{2t}. \tag{3-14}$$

其中

$$\begin{cases} \Pi_{10} = \dfrac{\gamma_0 - \beta_0}{\beta_1 - \gamma_1}, \Pi_{11} = \dfrac{\gamma_2}{\beta_1 - \gamma_1}, \\ \Pi_{20} = \dfrac{\beta_1\gamma_0 - \beta_0\gamma_1}{\beta_1 - \gamma_1}, \Pi_{21} = \dfrac{\beta_1\gamma_2}{\beta_1 - \gamma_1}. \end{cases} \tag{3-15}$$

在(3-15)式中，只有 4 个 Π 系数方程，将无法确定全部 5 个结构系数 β_0，β_1 和 γ_0，γ_1，γ_2．进一步的演算可知 $\beta_0 = \Pi_{20} - \beta_1\Pi_{10}$，$\beta_1 = \Pi_{21}/\Pi_{11}$．但 γ_0，γ_1 及 γ_2 则无法确定．这就是说，通过对诱导方程(3-13)和(3-14)的一致性估计，可得到需求方程(3-11)的一致估计，但却得不到供给方程(3-12)的一致估计．

是否能从模型的诱导式系数(Π)推出某一结构方程所含的全部参数，取决于该结构方程的**可识别性**(identifiability)．一个结构方程可识别，是指可通过模型的诱导式的一致性估计推算出该结构方程的一致性估计；否则，它就是不可识别(unidentifiable 或 underidentified)的．就上面的供求模型来说，需求方程是可识别的，而供给方程是不可识别的．

3.2 识别问题

一个结构方程可否识别,乃相对于其余方程的设定情况而言.如果对它的设定无法使它区别于其余方程,或区别于与其余方程相并的一个混合体,则是不可识别的.对不可识别的方程进行估计将是无意义的.因此,在考虑估计方法之前,应先考虑识别问题.

假定在工时的供给方程(3-2)中去掉其他收入和抚养儿童数两个外生变量,那么会出现什么问题呢?这时结构模型变为

需求: $工时_i = \beta_0 + \beta_1 工资_i + u_i$, (3-16)

供给: $工时_i = \gamma_0 + \gamma_1 工资_i + v_i$. (3-17)

其中只有两个内生变量,而无外生变量.它们的系统部分的图形以及对工资-工时的观测点如图 3-2 所示.根据拟合 LS 回归线的经验,也许认为总可以画出一条工时对工资的回归线.然而,谁能说这根线是需求线还是供给线,或者是什么别的线呢?看来可不能.这就是说无法对需求系数或供给系数作出有意义的估计,换言之,它们是不可识别的.

假定正确的供给方程包含其他收入一项、但不包含抚养儿童数,则整个模型变为(3-11)式和(3-12)式.这样,图形的系统部分如图 3-3 中更多的供给线段所示.供给线是诸多根平行线段中的哪一根,就决定于其他收入的大小了.然而,不难看出,如果按照拟合 LS 回归线的经验,利用图中随其他收入而变化的观测点,估计出来的回归线就应该是需求线.像这样的情形,就认为模型中的需求方程可以识别、而供给方程则不可识别.

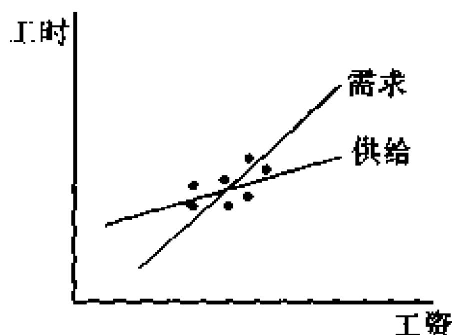


图 3-2 供需两线均不可识别

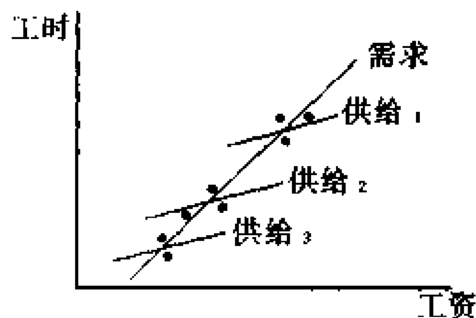


图 3-3 需求可识别,供给不可识别

可识别和不可识别的区分,关键在什么地方?比较一下(3-11)式和(3-12)式中的变量,便看出在需求方程中把其他收入这个外生变量排斥掉是识别的关键所在.

3.2.1 阶条件

如果联立方程模型中的每一个方程都可识别,就说这个模型是可识别的.要判断模型中某一个方程是否可识别,一般都要用联立方程的矩阵代数方法才便于导出它的识别条件.这里从略.

在一个现实的结构模型中,必定含有若干个内生变量和若干个**前定**(predetermined)(指外生和内生滞后)**变量**.一个方程能被识别的关键在于:并不是模型的全部变量都出现在这个方程之中.那些不出现的变量,内生也好,外生也好,都称之为被排斥的(excluded)变量.

识别的一个必要条件是:如果一个方程可以识别,则被排斥的前定变量的个数必须大于或等于它所含内生变量的个数减1.这个必要条件称为**阶条件**(order condition).

如果上述阶条件中的"大于"适用,则该方程是"**过度识别**"(overidentified)的;如果"等于"适用,则方程是"**恰好识别**"(just identified)的.虽然阶条件是必要的而不是充分的,但它在实践中被广泛地采用为识别的评判准则.之所以这样做是因为满足阶条件的方程很少是不可识别的,而要引用充分条件却是相当麻烦的.

现利用阶条件判别由(3-11)式和(3-12)式构成的模型.(3-11)式中被排斥的外生变量个数为1,内生变量的个数是2;按阶条件,$1=2-1$,故(3-11)式可识别.但(3-12)式不可识别(因为$0<2-1$).

如果一个方程不可识别,估计它就没有意义.例如,假定用了某种方法把(3-12)式估计出来了,但这并不能使人相信所估计的方程就是供给方程(3-12),因为如把(3-11)式和(3-12)式按某种方式相并,比如按各占50%的比例相加,便得

$$\text{工时}_i=\frac{\beta_0+\gamma_0}{2}+\frac{\beta_1+\gamma_1}{2}\text{工资}_i+\frac{\gamma_2}{2}\text{其他收入}_i+\frac{u_i+v_i}{2}$$

或

$$\text{工时}=\alpha_0+\alpha_1\text{工资}_i+\alpha_2\text{其他收入}_i+w_i. \tag{3-18}$$

其中,

$$\alpha_0=\frac{\beta_0+\gamma_0}{2},\alpha_1=\frac{\beta_1+\gamma_1}{2},\qquad \alpha_2=\frac{\gamma_2}{2},w_i=\frac{u_i+v_i}{2}.$$

那么,用同样的观测值去估计(3-12)式和(3-18)式会得到相同的结果,所估计的系数既可说是β,也可说是α.这当然就不能肯定所估计的是供给方程了.事实上(所估计的)(3-12)式或(3-18)式均可被解释为供需两方程的任意线性组合,它们在观测上是等效的.

还可以从诱导型系数和结构型系数之间的关系来说明这个识别和估计的关系问题.如果一个方程不是恰好识别的,就不能从诱导型系数唯一地解出结构型系数.当它是不可识别时,无解;而当它是过度识别时,有多个解;仅当它是恰好识别时,才有唯一解,这时才好用ILS法.

3.2.2 秩条件

如上所述,阶条件仅是可识别性的必要条件而非充分条件.例如需求方程(3-11)之所以满足阶条件,是因为它不含有供给方程(3-12)式所含的"其他收入"项.但可识别性的实现,还需要供给方程中"其他收入"项的系数γ_2不为零(也就是仅当"其他收入"项不仅有可能而且确实出现在供给方程之中)才能保证.

为了便于下面的叙述,令

M = 模型中内生变量的个数，

m = 某特定方程中内生变量的个数，

K = 模型中前定变量的个数，

k = 某特定方程中前定变量的个数.

某特定方程即使满足阶条件，即

$$K - k \geqslant m - 1,$$

仍有可能因为(出现在模型中但不为该方程所含有的)那些($K - k$ 个)被排斥的前定变量不全是独立的，致使结构型系数(β)和诱导型系数(Π)之间没有一一对应关系，而不可识别. 下面给出的秩条件则是可识别性的充分且必要条件.

秩条件(rank condition)：在含 M 个内生变量的 M 个联立方程模型中，某一方程是可识别的. 当且仅当模型所含而该方程所不含的变量(内生或前定)的系数矩阵能提供至少一个$(M-1)\times(M-1)$阶的非零行列式.

为了说明秩条件的应用，下面以如下的一个人为的($M=4, K=3$)联立方程组(3-19)式～(3-22)式来说明，其中 Y 和 X 分别代表内生和前定变量.

$$Y_{1t} - \beta_{10} - \beta_{12}Y_{2t} - \beta_{13}Y_{3t} - \gamma_{11}X_{1t} = u_{1t}, \quad (3\text{-}19)$$

$$Y_{2t} - \beta_{20} - \beta_{23}Y_{3t} - \gamma_{21}X_{1t} - \gamma_{22}X_{2t} = u_{2t}, \quad (3\text{-}20)$$

$$Y_{3t} - \beta_{30} - \beta_{31}Y_{1t} - \gamma_{31}X_{1t} - \gamma_{32}X_{2t} = u_{3t}, \quad (3\text{-}21)$$

$$Y_{4t} - \beta_{40} - \beta_{41}Y_{1t} - \beta_{42}Y_{2t} - \gamma_{43}X_{3t} = u_{4t}. \quad (3\text{-}22)$$

为了便于判断，可将方程组列成一个明显的表格形式，见表 3-1.

表 3-1

方程编号	变量的系数							
	1	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3
(3-19)	$-\beta_{10}$	1	$-\beta_{12}$	$-\beta_{13}$	0	$-\gamma_{11}$	0	0
(3-20)	$-\beta_{20}$	0	1	$-\beta_{23}$	0	$-\gamma_{21}$	$-\gamma_{22}$	0
(3-21)	$-\beta_{30}$	$-\beta_{31}$	0	1	0	$-\gamma_{31}$	$-\gamma_{32}$	0
(3-22)	$-\beta_{40}$	$-\beta_{41}$	$-\beta_{42}$	0	1	0	0	$-\gamma_{43}$

运用阶条件，容易判知每个方程都满足阶条件. 现再按秩条件作进一步的判断. 考虑到第一个方程(3-19)式，它不含变量 Y_4，X_2 和 X_3，为使该方程能被识别，必须能从该方程所不包含而又为其余方程所包含的变量的系数矩阵中，至少划出一个 3×3 阶的非零行列式来. 为此，先将对应于方程(3-19)的一行系数划掉，再将相对于(3-19)式来说是非零的列划掉. 把剩下的系数矩阵记为 $\boldsymbol{A}$，有

$$\boldsymbol{A} = \begin{bmatrix} 0 & -\gamma_{22} & 0 \\ 0 & -\gamma_{32} & 0 \\ 1 & 0 & -\gamma_{43} \end{bmatrix},$$

从 $\boldsymbol{A}$ 只能构造出唯一的一个 3×3 行列式

$$\det A = \begin{vmatrix} 0 & -\gamma_{22} & 0 \\ 0 & -\gamma_{32} & 0 \\ 1 & 0 & -\gamma_{43} \end{vmatrix}.$$

由于 $\det A = 0$，即 A 的秩小于3，因此方程不满足秩条件，从而判知它不可识别．$\det A = 0$ 表明 A 的行或列并非线性独立，这意味着变量 Y_4，X_2 和 X_3 之间存在某种数值关系，致使没有足够的信息能教人从已知的诱导型系数推算出结构型系数来．

一般地，常用秩条件判别一个方程是不是可识别的；如果可识别，则用阶条件进一步判别是恰好识别还是过度识别．

一个变量被排斥在一个方程之外，也可说成这个变量的系数被约束为零．除了利用这种零约束来判断可识别性外，还可借助于许多其他的约束形式，如对系数的线性或非线性约束，对误差项的约束等，来确定一个方程是否可以识别．

3.3　估计方法

联立方程的估计方法可分两类：一类是单一方程方法，又称**有限信息法**(limited information methods)；另一类是方程组方法(system methods)，又称**完全信息法**(full-information methods)．简单地说，前者就是对方程组中的每一个方程个别地进行估计，只考虑该方程所受的约束，而不问其余方程受到什么限制；后者则适当考虑方程组中每一道方程所受的约束，同时并举地估计全部方程．例如，设模型由(3-19)式～(3-22)式组成．如果要估计(3-21)式，单一方程法仅要求看到该方程不含 Y_2 和 X_3，而不问其他方程如何．方程组法则要兼顾其他方程所受的限制，如方程(3-19)中不含 Y_4，X_2 和 X_3，方程(3-20)和(3-22)中又不含什么，等等．当然，还可能有方程与方程之间(比方说，误差项与误差项之间)的约束条件，等等．虽然方程组法有信息方面的优越性，实际上却很少应用．其原因一方面是计算复杂(特别是大型的宏观计量经济模型动辄包含数百个方程)，另一方面是信息上的优越性是建立在模型设定的正确性基础之上的．一旦模型设定有误，哪怕是一道方程或一个约束条件有误，都会影响全部计算的效果．

基于上述考虑，下面将着重介绍单一方程法中的主要方法——二段最小二乘(two stage least squares)法．

3.3.1　二段最小二乘法

为了消除用OLS法直接估计结构方程的联立性偏误，方法之一是用ILS法．但当结构方程为过度识别情形时，ILS无法给出结构系数的唯一解．为克服这一困难，巴斯曼(R.L.Basmann)和泰尔(H.Theil)设计了**二段最小二乘(TSLS)法**．此法至今是人们估计联立方程模型最常用的方法．它的思想方法是，先用LS法把结构方程右边的每一个内生解释变量换成一个与误差项无关的替代变量，然后再用LS法去估计其结构系数．具体地说，假定整个方程组的全部前定变量是 $Z_1, Z_2, \cdots, Z_k$，再假定待估的结构方程除去左端有一个内生变量外，右端还含有内生解释变量 Y_1, Y_2，

$\cdots, Y_s$. 这时，需要用 LS 法求每个 $Y_j(j=1,2,\cdots,s)$ 对全部前定变量 $Z_1, Z_2, \cdots, Z_k$ 的回归方程. 记所求得的 LS 回归系数为 $\hat{\Pi}_{j0}, \hat{\Pi}_{j1}, \cdots, \hat{\Pi}_{jk}$，则 Y_j 的替代变量就是

$$\hat{Y}_j = \hat{\Pi}_{j0} + \hat{\Pi}_{j1} Z_1 + \cdots + \hat{\Pi}_{jk} Z_k. \tag{3-23}$$

可见，$\hat{Y}_j$ 是利用 Y_j 的诱导方程估算出来的.

假定模型由(3-1)和(3-2)两式构成，其中(3-1)式可识别. 现以(3-1)式为例，它的 TSLS 步骤是：

(1) 写出工资的诱导方程

$$\text{工资}_i = \Pi_0 + \Pi_1 \text{其他收入}_i + \Pi_2 \text{儿童数}_i + u_i,$$

用 LS 法估计诱导方程得

$$\widehat{\text{工资}}_i = \hat{\Pi}_0 + \hat{\Pi}_1 \text{其他收入}_i + \hat{\Pi}_2 \text{儿童数}_i. \tag{3-24}$$

(2) 将$\widehat{\text{工资}}_i$ 代入(3-1)式中的工资$_i$，因工资$_i$ = $\widehat{\text{工资}}_i + \hat{u}_i$，故

$$\text{工时}_i = \beta_0 + \beta_1 \widehat{\text{工资}}_i + u_i^*,$$

其中 $u_i^* = u_i + \beta_1 \hat{u}_i$. 再用 LS 法估计 β 系数.

在(3-23)式或(3-24)式中，由于前定变量 $Z_i(i=1,2,\cdots,k)$ 与全部误差项无关，$\hat{\Pi}$ 又是 Π 的一致性估计，从而 $\hat{Y}_j$ 可基本上视为诸 Z_i 的线性组合，故可认为 $\hat{Y}_j$ 与误差项渐近无关. 这样就说明 TSLS 法估计的一致性.

事实上，TSLS 法是一种工具变量法(参看 2.1.2 节). 前面说以 $\hat{Y}$ 作为 Y 的替代变量，其实是用 $\hat{Y}$ 作为工具变量. 为了易于说明，不妨用矩阵符号把模型的第一个结构方程写为

$$\boldsymbol{y} = \boldsymbol{Y}_1 \boldsymbol{\beta} + \boldsymbol{X}_1 \boldsymbol{\gamma} + \boldsymbol{u}.$$

其中 $\underset{n\times 1}{\boldsymbol{y}}$，$\underset{n\times(n-1)}{\boldsymbol{Y}_1}$ 是第一个结构方程所含的内生变量(非滞后)；$\underset{n\times k}{\boldsymbol{X}_1}$ 是它所含的前定变量. 第一段 LS 是求 $\boldsymbol{Y}_1$ 中每一内生变量对模型中的全部前定变量的 LS 回归，得

$$\hat{\boldsymbol{Y}}_1 = \boldsymbol{X}\hat{\boldsymbol{\Pi}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_1,$$

$\boldsymbol{X}$ 为模型的全部前定变量.

第二段 LS 是求 $\boldsymbol{y}$ 对 $\hat{\boldsymbol{Y}}_1$ 和 $\boldsymbol{X}_1$ 的 LS 回归，即

$$\begin{bmatrix} \hat{\boldsymbol{Y}}_1^{\mathrm{T}}\hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_1^{\mathrm{T}}\boldsymbol{X}_1 \\ \boldsymbol{X}_1^{\mathrm{T}}\hat{\boldsymbol{Y}}_1 & \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{Y}}_1^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}, \tag{3-25}$$

其中$\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}$，或记$\begin{bmatrix} \boldsymbol{b}_{\mathrm{TSLS}} \\ \boldsymbol{c}_{\mathrm{TSLS}} \end{bmatrix}$，是$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$的 TSLS 估计量.

若记 $\boldsymbol{Z} = [\boldsymbol{Y}_1 \quad \boldsymbol{X}_1]$，$\hat{\boldsymbol{Z}} = [\hat{\boldsymbol{Y}}_1 \quad \boldsymbol{X}_1]$ 和 $\boldsymbol{\alpha}^{\mathrm{T}} = [\boldsymbol{\beta}^{\mathrm{T}} \quad \boldsymbol{\gamma}^{\mathrm{T}}]$，则有

$$\hat{\boldsymbol{Z}}^{\mathrm{T}} \hat{\boldsymbol{Z}} \boldsymbol{\alpha} = \hat{\boldsymbol{Z}}^{\mathrm{T}} \boldsymbol{y},$$

然而,可以证明

$$\hat{Z}^{\mathrm{T}}\hat{Z}=\hat{Z}^{\mathrm{T}}Z,$$

故

$$\hat{Z}^{\mathrm{T}}\hat{Z}\alpha=\hat{Z}^{\mathrm{T}}Z\alpha=\hat{Z}^{\mathrm{T}}y,$$

于是得

$$\alpha=(\hat{Z}^{\mathrm{T}}Z)^{-1}\hat{Z}^{\mathrm{T}}y.$$

对照(2-13)式,便知 Z 是工具变量,可以证明,在一些标准的条件下,TSLS 估计量 $\alpha^{\mathrm{T}}=[\beta^{\mathrm{T}}\quad\gamma^{\mathrm{T}}]$的渐近方差

$$\text{asy var}\begin{bmatrix} b_{\text{TSLS}} \\ c_{\text{TSLS}} \end{bmatrix} = S^2\begin{bmatrix} \hat{Y}_1^{\mathrm{T}}\hat{Y}_1 & \hat{Y}_1^{\mathrm{T}}X_1 \\ X_1^{\mathrm{T}}\hat{Y}_1 & X_1^{\mathrm{T}}X_1 \end{bmatrix}^{-1}$$

$$= S^2\begin{bmatrix} Y_1X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_1 & Y_1^{\mathrm{T}}X_1 \\ X_1^{\mathrm{T}}Y_1 & X_1^{\mathrm{T}}X_1 \end{bmatrix}^{-1}.$$

其中

$$S^2=\frac{(y-Y_1b-X_1c)^{\mathrm{T}}(y-Y_1b-X_1c)}{n-m-k+1},$$

m = 方程所含内生变量个数,k = 方程所含前定变量个数.

注意:上式右端的内生解释变量是 Y_1 而不是 $\hat{Y}_1$.

需要指出,仅当识别的必要条件成立时,才能从正规方程(3-25)式解出$\begin{bmatrix} b \\ c \end{bmatrix}$,在恰好识别的情形下,TSLS 无异于 ILS.

究竟 OLS 和 TSLS 相差多大? 实际计算表明,有时 OLS 和 TSLS 的估计结果十分相近($\hat{Y}_1$ 和 Y_1 高度相关),但有时又差异较大.那么,到底有没有必要作 TSLS 估计? OLS 估计是否会比 TSLS 估计更好? 这些问题都是难于明确回答的.当多重共线性较强时,两种估计方法都有较大偏误.但当抽样波动较大,不宜只考虑估计方法上的偏误时,OLS 可能在某些情况下优于 TSLS.然而,更多的迹象表明,TSLS 是更为可取的方法.若计算条件允许,应尽可能使用,或者把 OLS 和 TSLS 的估计结果并列出来,让读者去评比、选择.

3.3.2　递归系统

有一种特殊的同时也是重要的联立方程模型,可以避免复杂的估计和识别问题,这就是所谓递归(recursive)系统.在这个系统中,可以把方程排成这样的顺序:前面出现的内生变量依赖于后来出现的内生变量,而后来的却不依赖于前面的.换句话说,因果关系是单向的.

例如,两个方程

$$Y=\beta_0+\beta_1X+u, \tag{3-26}$$

$$X=\gamma_0+\gamma_1Z+v, \tag{3-27}$$

其中，Y，X 为内生变量，Z 为外生变量. 方程(3-26)式表明，Y 依赖于 X，而方程(3-27)式表明 X 不依赖于 Y. 如再假定其中的误差项 u 和 v 不相关，就可看出 X 和 u 不相关，从而 OLS 适用于估计每一方程，而且没有识别问题.

相对于递归系统来说，平常说的联立方程模型大多属于一种**相依系统**(interdependent system). 在递归系统中内生变量之间的关系是单向的，而在相依系统中内生变量之间的关系是双向的.

为了把递归系统叙述得更为一般，可把它的形式适当扩充，如

$$
\begin{aligned}
Y_1 &= \beta_{10} + Y_2\beta_{12} + Y_3\beta_{13} + Z_1\gamma_1 + u_1,\\
Y_2 &= \beta_{20} + Y_3\beta_{23} \qquad\qquad + Z_2\gamma_2 + u_2,\\
Y_3 &= \beta_{30} \qquad\qquad\qquad\qquad + Z_3\gamma_3 + u_3,
\end{aligned}
$$

其中，Y_i 为内生变量，Z_j 为外生或前定变量.

$$\mathrm{cov}(u_1, u_2) = \mathrm{cov}(u_1, u_3) = \mathrm{cov}(u_2, u_3) = 0.$$

还可把 Y_i，Z_j 看做向量，并且每个 Y_i 所含元素全不相同，自然，β，γ 和 u 也都是相适(conformable)向量. 于是每一方程代表一个方程组，作为整个模型的一个分块，而整个模型就成为一个**分块递归**(block recursive)系统：前面分块的内生变量依赖于后面分块的内生变量，但后面的不依赖于前面的，这样就能把由联立性引起的估计和识别问题限于每个分块之内，分块之间便没有联立性所带来的麻烦的估计和识别问题了.

3.3.3 “外生性”检验

在联立方程模型中，外生变量和内生变量的区分，是估计和识别问题的关键所在. 传统的计量经济模型设计者，对外生和内生变量的分类，是凭先验知识作出判断的. 即使模型的设计者受到某种经济理论观点的指引，也会因观点的不同而对变量作出相异的分类. 为了提供一种客观的判别准则，下面介绍一种**“联立性”检验**，它是豪斯曼(J. A. Hausmann)模型设定检验的一个应用.

假定一个含有 3 个方程的联立方程模型，其中有 3 个内生变量 Y_1，Y_2 和 Y_3 及 3 个外生变量 X_1，X_2 和 X_3，并且第一个方程是

$$Y_1 = \beta_0 + \beta_2 Y_2 + \beta_3 Y_3 + \alpha_1 X_1 + u_1. \tag{3-28}$$

为了判断可否用 OLS 法估计(3-28)式，需确定 Y_2 和 Y_3 可否当作“外生”变量处理. 为此，可通过诱导方程(像做 TSLS 的第一段 LS 那样)，得到 Y_2 和 Y_3 的 LS 估计 $\widehat{Y}_2$ 和 $\widehat{Y}_3$，然后利用这两个估计量做如下的 LS 回归：

$$Y_1 = \beta_0 + \beta_2 Y_2 + \beta_3 Y_3 + \alpha_1 X_1 + \gamma_2 \widehat{Y}_2 + \gamma_3 \widehat{Y}_3 + u_1, \tag{3-29}$$

以检验假设 $H_0: \gamma_2 = \gamma_3 = 0$(可用 F 检验). 如果 H_0 被拒绝，就认为 Y_2 和 Y_3 不能被看做“外生”；反之，如果接受 H_0，就把 Y_2 和 Y_3 看做“外生”.

上述方法的逻辑性在于：通过对诱导式的估计可得 $Y_j = \widehat{Y}_j + \widehat{e}_j (j = 2,3)$，其中 $\widehat{e}_j$ 为诱导方程的 LS 误差估计. $\widehat{Y}_j$ 作为外生变量的线性组合(渐近意义的)，将与误

差 u_1 无关. 若将 $Y_j=\widehat{Y}_j+\widehat{e}_j$ 代入(3-28)式做如下的回归:

$$Y_1=\beta_0+\beta_2\widehat{Y}_2+\beta_3\widehat{Y}_3+\alpha_1X_1+\lambda_2\widehat{e}_2+\lambda_3\widehat{e}_3+u_1. \tag{3-30}$$

(其中 $\beta_j=\lambda_j, j=2,3$),并通过检验 $H_0:\lambda_2=\lambda_3=0$ 而认为 $\lambda_2=\lambda_3=0$,则(比较(3-28)式和(3-30)式)可将 Y_2 和 Y_3 视同 $\widehat{Y}_2$ 和 $\widehat{Y}_3$,亦即视同"外生". 可以证明,对(3-29)式检验 $H_0:\gamma_2=\gamma_3=0$,渐近等效于对(3-30)式检验 $H_0:\lambda_2=\lambda_3=0$(其中 $\gamma_j=\beta_j=\lambda_j$),但通过(3-29)式做检验更为有效.

3.4　模拟与预测

一个涉及不同时期的内生变量的结构模型如3.1节例2,具有描述每一个内生变量的时间走道的功能,可用于预测任一内生变量的未来情形. 为此,可通过解联立线性方程组的方法,将其结构方程转换为诱导方程:

$$\begin{cases} Y_t=\Pi_{10}+\Pi_{11}g_t+\Pi_{12}r_t+\Pi_{13}r_{t-1}+\Pi_{14}Y_{t-1}+\Pi_{15}Y_{t-2}+\varepsilon_{1t}, \\ C_t=\Pi_{20}+\Pi_{21}g_t+\Pi_{22}r_t+\Pi_{23}r_{t-1}+\Pi_{24}Y_{t-1}+\Pi_{25}Y_{t-2}+\varepsilon_{2t}, \\ i_t=\Pi_{30}+\Pi_{31}g_t+\Pi_{32}r_t+\Pi_{33}r_{t-1}+\Pi_{34}Y_{t-1}+\Pi_{35}Y_{t-2}+\varepsilon_{3t}. \end{cases} \tag{3-31}$$

其中,$\Pi_{10}=(\alpha_0+\beta_0)\delta, \Pi_{11}=\delta, \Pi_{12}=\alpha_2\delta, \Pi_{13}=\beta_2\delta, \Pi_{14}=-\Pi_{15}=\beta_1\delta, \delta=1/(1-\alpha_1(1-\tau))$,余类推;$\varepsilon_{1t},\varepsilon_{2t},\varepsilon_{3t}$ 皆是 u_t 和 v_t 的线性组合.

从预测的角度考虑,在诱导方程中,除左端一个内生变量外,右端的变量都是前定变量,故从右端预测左端,可以说是"真实"的预测. 对比之下,在结构方程中,因涉及从右端的内生变量到左端的内生变量的推算,而右端的内生变量又不能事先给定,故属于一种理论上的"预测".

凡是已估算出来的结构方程(指其结构系数已被估算出来),不论用什么估算方法,均称**模拟方程**. 从模拟方程组解出内生变量,无异于从结构方程解出诱导方程. 但这时诱导方程的系数有了具体数值(从结构式的系数值推算诱导式的系数值总是可能的). 于是,对给定的外生变量值,并令干扰项=0,即可对解出来的内生变量进行预测. 若把外生变量看做政策变量,通过预测便可进行政策性分析.

但是,诱导方程主要适用于短期预测. 因为,一般来说,一个诱导方程除含有外生变量外,还会含有不同内生变量的滞后项. 例如消费 C 的诱导式中含有 Y 的滞后项,为了分析 C 如何在外生变量的影响下发生变化,不免还掺杂着最近、过去的 Y 的影响,而不能纯粹地看到外生变量对 C 的连续影响. 为了描述外生变量的某种时间走道如何影响 C 的时间走道,还需要把诱导方程转换成除含有外生变量外仅含 C 这一个内生变量及其滞后项的形式. 现在,Y 的诱导式恰好是这种形式,即除外生变量外,只 Y 本身及其滞后项. 下面就拿它来做模拟试验和预测分析.

例如,若想知道如果政府从 T 时期起增加100亿元的开支会对国民收入产生什么影响,不妨把(3-31)式写成差分形式. 记 $\Delta Y_t=Y_t-Y_{t-1}$,可得

$$\Delta Y_t=\Pi_{11}\Delta g_t+\Pi_{12}\Delta r_t+\Pi_{13}\Delta r_{t-1}+\Pi_{14}\Delta Y_{t-1}+\Pi_{15}\Delta Y_{t-2}. \tag{3-32}$$

为了便于说明,固定 r 不变,即 $\Delta r_t=\Delta r_{t-1}=0$,则(3-32)式就简化为

$$\Delta Y_t = \Pi_{11}\Delta g_t + \Pi_{14}\Delta Y_{t-1} + \Pi_{15}\Delta Y_{t-2}. \quad (3\text{-}33)$$

设政府在时期 $t+1$ 增加开支 $d=100$ 亿元，且以后一直保持这一新的开支水平，即

$$\Delta g_{t+1} = d, \quad \Delta g_{t+2} = \Delta g_{t+3} = \cdots = 0. \quad (3\text{-}34)$$

(3-34)式代表外生变量 g 的一种时间走道. 设想 g 和 r 已在过去足够长的一段时间里保持不变，以致 Y 能稳定在某个不变的均衡水平上，即 $\Delta Y_t = \Delta Y_{t-1} = \Delta Y_{t-2} = \cdots = 0$，那么，由(3-33)式知，$d$ 对国民收入的倍数效应是

在时期 $t+1$，$\Delta Y_{t+1} = \Pi_{11} d$，

在时期 $t+2$，$\Delta Y_{t+2} = \Pi_{14}\Delta Y_{t+1} + \Pi_{15}\Delta Y_t = \Pi_{14}\Pi_{11} d$，

在时期 $t+3$，$\Delta Y_{t+3} = \Pi_{14}\Delta Y_{t+2} + \Pi_{15}\Delta Y_{t+1} = (\Pi_{14}^2\Pi_{11} + \Pi_{15}\Pi_{11}) d$，

…

由此得

即期倍数 Π_{11}；

中期倍数 $\Pi_{14}\Pi_{11}, \Pi_{14}^2\Pi_{11} + \Pi_{15}\Pi_{11}, \cdots$；

长期倍数 $\Pi_{11} + \Pi_{14}\Pi_{11} + \Pi_{14}^2\Pi_{11} + \Pi_{15}\Pi_{11} + \cdots$.

这些倍数说明了增加政府开支 $d=100$ 亿元对国民收入 Y_t 在不同时期的影响.

方程(3-31)式除外生变量外，只含一个内生变量 Y 的过去和现在. 这样的方程称为**基本动态方程**(fundamental dynamic equation). 为了了解外生变量对某一内生变量的倍数效应，就需要导出相应于这个内生变量的基本动态方程.

把(3-31)式写成

$$Y_t - \Pi_{14}Y_{t-1} - \Pi_{15}Y_{t-2} = f(g,r) + \varepsilon_{1t},$$

其中 $f(g,r) = \Pi_{10} + \Pi_{11}g_t + \Pi_{12}r_t + \Pi_{13}r_{t-1}$. 这是一个二阶差分方程，其自由项为 $f(g,r)$. 可以证明，对线性联立方程模型来说，每一内生变量的基本动态方程都有同样的齐次部分(指方程左边的 Π 系数)，所不同的仅是自由项. 例如，投资的基本动态方程是

$$i_t - \Pi_{14}i_{t-1} - \Pi_{15}i_{t-2} = h(g,r) + \text{误差},$$

对消费也是如此. 所以它们都有共同的动态特征.

还可以通过对参数的适当改动，做其他方面的政策分析. 例如，根据理论分析或实际计算，如预期某项减税措施会起到提高边际消费倾向 $\alpha\%$ 的效果，就可作减税前和减税后两个模拟试验以资比较，看结果是否和预期的相一致.

4 定性与限值应变量

在统计调查中，常遇到**定性**(qualitative)或限值应变量(limited dependent variables)问题. 如某家庭是否购买某一商品；某人是否接受某一职位；某家庭对耐用消费品的开支要么是一大笔钱，要么等于零，而不是一个连续变量. 又如，某一戏院的门票销售量受到戏院座位的限制，一些商品的买卖有时受到限量和限价的约束，等等. 从 20 世纪 60 年代前后开始，随着家计调查数据的广泛收集，如何建立定性或

限值应变量模型,对人们的这些买卖或选择行为作出解释,已成为近代计量经济学发展的一个活跃的研究领域.

4.1　线性概率模型

设想一个家庭买或不买一件耐用消费品如洗衣机,与它的收入有关.买或不买可由一个虚拟变量(即(0,1)变量)来表示.现考虑一个简单的回归模型

$$Y_i = \alpha + \beta X_i + u_i, \quad i = 1, \cdots, n, \tag{4-1}$$

其中,X = 家庭收入;

$$Y = \begin{cases} 1 & (\text{若购买洗衣机}), \\ 0 & (\text{若不购买}); \end{cases}$$

u = 随机干扰.

如假定　$E(u_i) = 0$,便有条件期望

$$E(Y_i | X_i) = \alpha + \beta X_i.$$

这个期望值可以解释为收入等于 X_i 的家庭购买洗衣机的概率.假定买和不买的概率分别为

$$P(Y_i = 1) = \Pi_i \quad 和 \quad P(Y_i = 0) = 1 - \Pi_i,$$

就知 Y_i 的期望值

$$E(Y_i) = 1 \cdot \Pi_i + 0 \cdot (1 - \Pi_i) = \Pi_i.$$

模型(4-1)也因此称为线性概率模型(linear probability model).其中自变量并不限于1个,可将 X_i 推广为向量,而将 $\alpha + \beta X_i$ 写为 $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i$.显然,在此模型中,

$$0 \leqslant E(Y_i | X_i) \leqslant 1.$$

在 α 和 β 的估计问题上,普通最小二乘法将会遇到一些困难:

(1) u_i 不是正态分布.

因 $u_i = Y_i - \alpha - \beta X_i$,故它只取两个值:

当 $Y_i = 1$ 时,$u_i = 1 - \alpha - \beta X_i$;

当 $Y_i = 0$ 时,$u_i = -\alpha - \beta X_i$.

(2) u_i 有异方差性.

为了使 $E(u_i) = 0$,u_i 的概率必须是

当 u_i 为 $-\alpha - \beta X_i$ 时,其概率为 $1 - \alpha - \beta X_i = 1 - \Pi_i$;

当 u_i 为 $1 - \alpha - \beta X_i$ 时,其概率为 $\alpha + \beta X_i = \Pi_i$.

因此,

$$\begin{aligned} \operatorname{var}(u_i) &= E(u_i^2) \\ &= (-\alpha - \beta X_i)^2 (1 - \alpha - \beta X_i) + (1 - \alpha - \beta X_i)^2 (\alpha + \beta X_i) \\ &= \Pi_i (1 - \Pi_i), \end{aligned}$$

将随 X_i 而变.

(3) 在实际估计中不能保证 $\Pi_i = E(Y_i | X_i)$ 的估计

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

确实落入区间[0,1]. 这是估计的真正困难所在. 如果用 OLS 法估计线性概率模型，不仅会出现 $\hat{Y}_i < 0$ 或 $\hat{Y}_i > 1$ 的情形，而且拟合适度低，判定系数大多在 0.2～0.6 之间.

4.2　概率单位与对数单位

解决线性概率模型估计问题的较好办法，是把 $\alpha + \beta X_i$ 换成某种概率单位. 不妨把 $\alpha + \beta X_i$ 写成更一般的向量形式 $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i$，以包括多个解释变量的情形. 设对应于 $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i$，$Y_i = 1$ 的概率为

$$P(Y_i = 1) = \Pi_i = F(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i),$$

要求变换 F 有以下两个性质是自然的，即

$$\underset{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i \to +\infty}{F} (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i) \to 1,$$

$$\underset{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i \to -\infty}{F} (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i) \to 0.$$

而这正好是任何适当的连续的累积概率分布函数的应有性质. 故作为变换 F，原则上可选用任何概率分布函数. 但在实际应用中，大多选用正态分布或者逻辑斯蒂(logistic)分布函数. 若假定 F 为累积正态分布函数，即

$$F(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i) = \Phi(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i) = \int_{-\infty}^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{t^2}{\alpha}\right) \mathrm{d}t. \tag{4-2}$$

这种变换称为**概率单位**(probit)或**正态单位**(normit)模型. 但它不是一个封闭形式而不便于分析运算. 现假定 $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i$ 的分布曲线是**逻辑斯蒂曲线**，即

$$\Pi_i = \frac{\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i}}{1 + \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i}} = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i}}. \tag{4-3}$$

这既是一个封闭而又接近正态的形式，而且不论 $\boldsymbol{X}$ 怎样变化，均能保证 $0 < \Pi_i < 1$.

于是，容易推出

$$\ln\left(\frac{\Pi_i}{1 - \Pi_i}\right) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i, \tag{4-4}$$

即 Π_i 和 $1 - \Pi_i$ 这两个概率之比的对数，或称**对数单位**(logit)，正好是 $\boldsymbol{X}_i$ 和 $\boldsymbol{\beta}$ 的线性函数. 变换(4-3)式或函数形式(4-4)称为对数单位模型.

4.2.1　广义最小二乘法

如果存在对 $\boldsymbol{X}_i$ 重复观测的数据，便可对对数单位模型做 GLS 回归估计如下：

为了估计 $\boldsymbol{\beta}$，将(4-4)式写为

$$\ln\left(\frac{\Pi_i}{1 - \Pi_i}\right) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i + u_i. \tag{4-5}$$

假定对相同的 X_i 作过 n_i 次重复观测,其中有 r_i 次 $Y_i=1$,就可把样本比值 $p_i=\frac{r_i}{n_i}$ 作为总体比值 Π_i 的近似值.可以证明,当 n_i(对不同的 i)均较大且每次观测结果可视同二项式变量那样独立地分布时,u_i 渐近于正态分布,即

$$u_i \sim N\left(0, \frac{1}{n_i \Pi_i(1-\Pi_i)}\right). \tag{4-6}$$

于是,在估计中用 p_i 代替(4-5)式和(4-6)式中的 Π_i,对(4-5)式做 GLS 回归估计,可以收到大样本渐近有效估计之效.

概括起来,如果存在对 X 适当划分的 X_i 的较多重复的观测数据,较好的估计方法将是

(1) 从样本比率 p_i 计算样本对数单位 $\ln\left(\frac{p_i}{1-p_i}\right)$.

(2) 用 p_i 代替 $(n_i\Pi_i(1-\Pi_i))^{-1}$ 中的 Π_i,用 $(n_ip_i(1-p_i))^{1/2}$ 通乘(4-5)式,然后对 β 做 LS 估计.

类似地,可对 probit 模型做 GLS 估计.所不同的仅是按照累积正态概率分布作变换.例如,$\boldsymbol{\beta}^{\mathrm{T}}=[\beta_1\ \beta_2]$,$X_i^{\mathrm{T}}=[1\ X_i]$.先按下面的关系式把 Π_i 换算成所谓效用指数(utility index) I_i:

$$I_i=\Phi^{-1}(\Pi_i)=\beta_1+\beta_2X_i.$$

由于 p_i 是 Π_i 的样本估计值,用 p_i 代替 Π_i 时上式可写为

$$\hat{I}_i=\Phi^{-1}(p_i)=\beta_1+\beta_2X_i.$$

然后利用实测的 X_i 和所换算的 $\hat{I}_i$ 对 $\hat{I}_i=\beta_1+\beta_2X_i+u_i$ 做 GLS 估计.

由于

$$\sigma_u^2=\frac{\Pi_i(1-\Pi_i)}{n_i\phi_i^2},$$

其中 ϕ_i 是对应于 $\Phi^{-1}(p_i)$ 的 ϕ 值,应用时取其样本估计 $\hat{\sigma}_u^2=\frac{p_i(1-p_i)}{n_i\phi_i^2}$.

4.2.2 最大似然法

如果缺少适当分组的重复观测数据(这是较常见的情形),则宜采用最大似然(ML)法.把每次观测视同伯努利(D. Bernoulli)试验(即二项式分布中的一次抽样),其成功概率为 $F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})$,$n$ 次独立观测的联合概率分布 P 或似然函数 L 为

$$P(Y_1=Y_1,Y_2=Y_2,\cdots,Y_n=Y_n)=\prod_{Y_i=0}(1-F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}))\prod_{Y_i=1}F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}),$$

或

$$L=\prod_{i=1}^{n}[F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)]^{Y_i}[1-F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)]^{1-Y_i},$$

$$Y_i=\begin{cases}1 & (\text{若成功(购买)});\\ 0 & (\text{若不成功}).\end{cases}$$

取对数得

$$\ln L=\sum_{i=1}^{n}\{Y_i\ln F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)+(1-Y_i)\ln[1-F(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)]\}.$$

其最大化的一阶条件为

$$\frac{\partial\ln L}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\left[\frac{Y_i f_i}{F_i}+(1-Y_i)\frac{-f_i}{1-F_i}\right]X_i=0.$$

其中 f 是 F 的导函数，下标 i 表明函数依赖于变量 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i$.

最大似然估计量及其信息矩阵的计算是比较麻烦的(常用迭代法). 但现在可利用计算机程序软件，如 SAS，RATS 等来计算，还包括对异方差性进行校正.

4.2.3 边际效应分析

一个在分析上要注意的问题是，尽管 logit 和 probit 模型的计算结果会给出很不相同的 β 估计值. 但从逻辑斯蒂曲线和正态曲线的近似程度看，两者的斜率系数(即 X 的变化对 Π 的边际效应)会是差不多的. 不妨拿这两个模型同线性概率模型(LPM)比较一下：

$$\text{LPM}: \Pi_i=\boldsymbol{\beta}^{\mathrm{T}}X_i,\quad \frac{\mathrm{d}\Pi_i}{\mathrm{d}\boldsymbol{X}_i}=\boldsymbol{\beta}.$$

$$\text{logit}: \Pi_i=(1+\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}X_i})^{-1},\quad \frac{\mathrm{d}\Pi_i}{\mathrm{d}\boldsymbol{X}_i}=\boldsymbol{\beta}\Pi_i(1-\Pi_i).$$

$$\text{probit}: \Pi_i=\Phi(\boldsymbol{\beta}\boldsymbol{X}_i),\quad \frac{\mathrm{d}\Pi_i}{\mathrm{d}\boldsymbol{X}_i}=\boldsymbol{\beta}\phi(\boldsymbol{\beta}\boldsymbol{X}_i).$$

为了分析边际效应，对 logit 模型来说，要计算 $\boldsymbol{\beta}\Pi_i(1-\Pi_i)$，实际上由 $\boldsymbol{\beta}p_i(1-p_i)$来替代；对 probit 模型来说，则要计算 $\boldsymbol{\beta}\phi(\boldsymbol{\beta}\boldsymbol{X}_i)$；但对 LPM 来说，$\boldsymbol{\beta}$ 就是它的斜率系数(或边际效应).

至于怎样选择模型，还不能单凭所估计 $\hat{\boldsymbol{\beta}}$的方差($\sigma_{\hat{\beta}}^2$)大小作为标准. 由于概率变换形式的选择有一定任意性，故还要看实际估计或预测的效果.

4.3 截取回归与断尾回归

假定要研究消费者打算花在购买房子的支出和他的收入(以及其他经济变量)之间的关系，那么会遇到一个难题：如果消费者不买房子，则显然不会有他为买房子而支出的任何数据；已经获得的购房数据仅限于实际上买了房子的消费者. 为此，可把消费者划分为两类：一类是兼能提供**回归元**(regressors)(如收入、抵押利率等)和**回归值**(regressand)(如购房开支)两方面信息的消费者；另一类是仅能提供回归元信息而不能提供回归值信息的消费者. 下面介绍解决这类问题的方法.

4.3.1 截取与断尾变量

设对大小为 n 的样本($Y_1^*,Y_2^*,\cdots,Y_n^*$)，仅如实记录大于 C 的 Y^* 值，对于小

于或等于 C 的 Y^* 值则一律记为 C.这样,从记录得到的观测值将是

$$Y_i = Y_i^*, \quad 若\ Y_i^* > C,$$
$$Y_i = C, \quad 若\ Y_i^* \leqslant C.$$

样本 $Y_1, Y_2, \cdots$ 称为截取样本.相应的被观测变量 Y 称为截取变量,其概率分布称为截取分布.显然,

$$Y_i = C \Rightarrow Y_i^* \leqslant C.$$

设 $Y^* \sim N(\mu, \sigma^2)$,则对给定的 Y,用以估计 μ 和 σ^2 的似然函数是

$$L(\mu, \sigma^2 \mid Y) = \prod_{Y^* > C} \frac{1}{\sigma}\phi\left(\frac{Y_i - \mu}{\sigma}\right) \prod_{Y^* \leqslant C} \Phi\left(\frac{C - \mu}{\sigma}\right).$$

其中 Φ 是正态(累积)分布函数,ϕ 是相应的密度函数.

若对 Y^* 的观测只限于大到(或小到)$Y^* = C$ 为止,而对 $Y^* > C$(或 $Y^* < C$)的 Y^* 根本不作记录,这样也就不知道有多少个 $Y^* > C$(或 $Y^* < C$).这样得到的样本 $Y^*(-\infty < Y^* \leqslant C)$(或 $Y^*(C \leqslant Y^* < +\infty)$)称为断尾样本.可相应地定义断尾变量和断尾分布.例如,断尾正态分布的密度函数为

$$f(Y^* \mid Y^* \leqslant C) = \frac{1}{\sigma}\phi\left(\frac{Y^* - \mu}{\sigma}\right) \Big/ \Phi\left(\frac{C - \mu}{\sigma}\right), \quad (-\infty < Y^* \leqslant C).$$

4.3.2 截取正态变量的均值与方差

下面仅叙述而不去证明一个关于截取正态变量的矩定理.

若 $Y^* \sim N(\mu, \sigma^2)$,并且

$$Y = C, \quad 若\ Y^* = C,$$
$$Y = Y^*, \quad 若\ Y^* > C,$$

则

$$E(Y) = Pr(Y = C) \times E(Y \mid Y = C) + Pr(Y > C) \times E[Y \mid Y > C]$$
$$= \Phi C + (1 - \Phi)(\mu + \sigma\lambda),$$
$$\mathrm{var}(Y) = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi].$$

其中:$\Phi[(C - \mu)/\sigma] = \Phi(\alpha) = Pr(Y^* \leqslant C) = \Phi, \lambda = \phi/(1 - \Phi)$,而 $\delta = \lambda^2 - \lambda\alpha$.
这是从下截取的情形.如果是从上截取,则要把式中 Φ 和 $(1 - \Phi)$ 对调,并且取 $\lambda = -\phi/\Phi$.见参考文献[3]或[4].

例1 现要研究人们对某戏院的某次演出的门票需求量.仅有的观测数据是实际售出的门票数.当一场演出的门票被售光时,可知道实际需求量大于出售额.即当需求量被转换成出售数时,该需求量已被截取为戏院的座位数.

假定戏院有 20 000 个座位,在当前季节里有 25% 的机会出现满座(即全部门票被售完).如果包括满座的情形在内,每场平均观众人数是 18 000 人,问门票需求的均值和方差是多少?

根据上述定理,18 000 将是如下数学期望的一个估计值:

$$E(销售量\ Y) = 20\,000(1 - \Phi) + (\mu + \sigma\lambda)\Phi.$$

因为本例是从上而不是从下截取,

$$\lambda = -\phi(\alpha)/\Phi(\alpha).$$

其中 $\alpha=(20\,000-\mu)/\sigma$. 因有 25% 的机会满座，故 $\Phi=0.75$. 代入逆标准正态函数得 $\alpha=\Phi^{-1}=0.675$，从而 $-\phi(0.675)/0.75=\lambda=-0.424$.

于是得到 μ 和 σ 的两个方程：

$$18\,000=0.25(20\,000)+0.75(\mu-0.424\sigma)$$

和

$$0.675\sigma=20\,000-\mu.$$

由此解出

$$\sigma=2\,426 \quad 和 \quad \mu=18\,362.$$

这里 μ 代表 E(潜在销售量 Y^*)的估计量.

作为比较，如果上述均值 18 000 仅代表非满座情况下的均值，但满座的情形占 25%. 那么，用以求解 μ 和 σ 的两个方程将是

$$18\,000=\mu-0.424\sigma$$

和

$$0.675\sigma=20\,000-\mu.$$

于是解得

$$\sigma=1\,820 \quad 和 \quad \mu=18\,772.$$

4.3.3 截取回归

这里研究如何把截取或断尾分布的参数估计和回归分析联系起来. 例如，研究电冰箱的需求量时，统计中人们填报用于购买电冰箱的费用 Y，要么大于 Y_0(设 Y_0 基本上代表电冰箱最低价格)，要么等于零. 因此在线性模型假设下，电冰箱开支 Y 和解释变量 $\boldsymbol{X}$ 之间的关系为

$$\begin{aligned} &Y=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+u, && 若\ Y\geqslant Y_0;\\ &Y=0, && 若\ Y<Y_0. \end{aligned}$$

即当 $Y<Y_0$ 时一律截取为零. 这就是**截取回归**(censored regression).

托宾(J. Tobin)1958 年提出了一种截取回归模型：

$$\begin{aligned} &Y_i=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i+u_i && 若等式右端>0, \quad i=1,\cdots,n.\\ &Y_i=0, && 若等式右端\leqslant 0. \end{aligned}$$

其中 $\boldsymbol{\beta}$ 是 $(k\times 1)$ 未知参数向量，$\boldsymbol{X}_i$ 是 $(k\times 1)$ 已知常数向量，$u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$. 该模型后来被称为**托宾模型**. 显然，托宾模型是截取回归模型的一种特殊情形.

托宾模型又可表示为

$$\begin{aligned} &Y_i^*=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i+u_i,\\ &Y_i=Y_i^*, \quad 若\ Y_i^*>0,\\ &Y_i=0, \quad\ \ 若\ Y_i\leqslant 0. \end{aligned}$$

这里 Y^* 扮演着一个潜变量的角色. 仅当它大于 0 时才是可观测的. 若将上述模型中第二个等式 $Y_i=Y^*$ 改为 $Y_i=1$，就成为 probit 模型. 因此**托宾模型**可看做是 probit 模型的延伸.

托宾模型还可表示为

$$Y_i = \max(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i + u_i, 0).$$

托宾模型要求在对 Y_i 和 $\boldsymbol{X}_i$ 的 $n(n>k)$ 次观测的基础上估计 $\boldsymbol{\beta}$ 和 σ^2. 对此,托宾指出,如果运用最小二乘法(LS)来估计,则困难较大;运用 OLS 估计,既有偏误,又具有非一致性. 因此,可行的是应用最大似然法来估计. 一些软件如 RATS,TSP,SAS 等已有计算托宾回归模型的 ML 程序,这里就不再赘述.

对托宾模型来说,潜在有三种条件均值可以考虑:一种是对潜变量的 $E(Y^*|\boldsymbol{X})$,它就是 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}$;另一种是对观测到的、包括截取和非截取的变量的 $E(Y|\boldsymbol{X})$,利用截取变量矩定理,有

$$E(Y|X) = \Phi C + (1-\Phi)(\mu + \sigma\lambda), \quad \lambda = \lambda(\alpha);$$

第三种是仅包括非截取观测值的变量的

$$E(Y|X; Y>0) = \mu + \sigma\lambda.$$

可以证明,$E(Y^*) < E(Y) < E(Y|Y>0)$. 注意,对不同条件的均值,边际效应是不同的. 例如,

$$\frac{\partial E(Y^*|X)}{\partial \boldsymbol{X}} = \boldsymbol{\beta},$$

而

$$\frac{\partial E(Y|X)}{\partial \boldsymbol{X}} = \boldsymbol{\beta}\Phi\left(\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}}{\sigma}\right).$$

在实际应用中,应考虑哪一种均值,将视研究的目的而定.

4.3.4 断尾回归

在回归中,如仅当应变量小于(或大于)等于某个门限值时,才具备有解释变量的观测值,则此时所做的回归称**断尾回归**. 例如,为实行某种补贴政策而调查居民收入,在这种调查中只收集有收入小于某门限值 C_i(C_i 是与居民 i 的特征有关的数)的样本信息. 若建立收入的回归方程,则对第 i 次观测可写成

$$Y_i = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i + u_i,\ u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2).$$

其中 $\boldsymbol{X}_i$ 代表向量解释变量如教育、年龄、经验等,收入 $Y_i \leqslant C_i$(C_i 是与家庭人口有关的常数)这一条件意味着

$$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i + u_i \leqslant C_i \quad 或 \quad u_i \leqslant C_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i.$$

很明显,数学期望 $E(u_i|u_i \leqslant C_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i) \neq 0$,而且它将是 $\boldsymbol{X}_i$ 的函数. 即是说,误差 u_i 与解释变量 $\boldsymbol{X}_i$ 相关,故普通最小二乘估计既是有偏误的,又是非一致性的.

断尾样本的似然函数是

$$L = \prod_{i=1}^{n} \frac{(1/\sigma)\phi[(Y_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)/\sigma]}{\Phi[C_i - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)/\sigma]}.$$

其中 Φ,ϕ 分别是标准正态分布与密度函数,n 代表样本大小. 可用牛顿-拉夫生迭代法求 β 和 σ^2 的最大似然估计.

4.4　非均衡模型

本节描述在供求不平衡的情况下，由超需(excess demand)或超供(excess supply)来刻画市场或计划经济过程的模型，目的在于获得供、求(及其调节)方程的估计．**非均衡模型**(disequilibrium models)的正确统计分析有赖于限值应变量的适当应用．

一个简单的非均衡模型有

需求函数　$D_t = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_t + u_t$，

供给函数　$S_t = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}_t + v_t$，

成交量　$Q_t = \min(D_t, S_t)$，

比例假设　$\Delta P_t = \gamma(D_t - S_t), \gamma > 0$．

其中误差项 u_t 和 v_t 既与序列无关，又无同期相关，并分别服从 $N(0, \sigma_u^2)$ 和 $N(0, \sigma_v^2)$．

模型的需求和供给函数可按通常方式详细列出，其中 $\boldsymbol{\alpha}, \boldsymbol{\beta}$ 为回归系数向量，解释变量 $\boldsymbol{X}_t$ 和 $\boldsymbol{Z}_t$ 可以包括价格 P_t，但 P_t 仍是内生变量．

最小条件 $Q_t = \min(D_t, S_t)$ 是一切非均衡模型的基石，用以代替均衡条件 $D_t = S_t$．最小条件表明，当供不应求即出现超需时，$Q_t = S_t$(成交量就是供给量)；当供过于求即出现超供(或负超需)时，$Q_t = D_t$(成交量就是需求量)．但在不能分清市场是超需还是超供的情况下，参数估计是非常困难的．为了分清 S_t 和 D_t，方法之一是利用比例假设．

可通过涨价期 $\Delta P_t = P_t - P_{t-1} > 0$ 和跌价期 $\Delta P_t < 0$ 来判别正的或负的超需，而将模型重新建立成限值应变量的形式．再由比例假设，当 $\Delta P_t > 0$ 时，

$$Q_t = S_t.\ D_t = S_t + (D_t - S_t) = Q_t + \Delta P_t / \gamma.$$

于是有需求函数

$$Q_t = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_t - \frac{1}{\gamma} \Delta P_t + u_t, \quad \Delta P_t > 0.$$

同理有供给函数

$$Q_t = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}_t + \frac{1}{\gamma} \Delta P_t + v_t, \quad \Delta P_t < 0.$$

将 ΔP_t 表示为潜变量，则模型又可写为

$$Q_t = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_t - \frac{1}{\gamma} g_t + u_t,$$

其中

$$g_t = \begin{cases} \Delta P_t, & \text{若 } \Delta P_t > 0; \\ 0, & \text{若不然}, \end{cases}$$

及

$$Q_t = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}_t - \frac{1}{\gamma} h_t + v_t,$$

其中

$$h_t=\begin{cases}-\Delta P_t, & 若\ \Delta P_t<0;\\ 0, & 若不然.\end{cases}$$

可采取适当的最大似然法或二段最小二乘法估计模型中的参数 $\alpha,\beta,r,\sigma_u^2$ 和 σ_v^2.

非均衡模型可用来描述计划经济如何根据超需或超供情况调节它的产量或价格的时间走道,或描述一个市场(如受价格管制的石油市场)的超需向另一市场(如有高度代替性的煤市场)溢出的多市场非均衡效应,等等.详见参考文献[6].

5 时间序列计量经济学方法

本章考虑时间序列的计量经济学方法问题.相对地说,前面所考虑的主要是频率分布领域的计量经济学方法,其中假定了每次"观测"都在"同一"的观测环境中进行,即有一个"不变"的频率或概率分布作为条件.把这个"不变"的条件移植到时间领域,就表现为时间序列的"**平稳性**".如果时间序列是不平稳的,那么,通常讲的以 t、F 或 χ^2 等检验为根据的假设检验方法,都会变成无效,以致造成"**谬误回归**"(spurious regression).

然而,经验表明,许多宏观经济变量的时间序列是不平稳的,那么,一个经济序列对另一个或一些经济序列的回归分析就需要有一定的合理程序.下面主要讨论:①如何检验一个时间序列的平稳性,以及如何化不平稳序列为平稳序列;②为了判断一个不平稳时间序列的回归分析是否有效,特别是提防"谬误"回归的出现,专门引入了"**协积**"(或"**协整**"(cointegration))的概念以及"**误差纠正机制**"(error correction mechanism)的应用.

5.1 趋势平稳与差分平稳

考虑时间序列 Y_t.如果

均值 $E(Y_t)=\mu$,

方差 $\mathrm{var}(Y_t)=E[(Y_t-\mu)^2]=\sigma^2$,

协方差 $\mathrm{cov}(Y_t,Y_{t-k})=E[(Y_t-\mu)(Y_{t-k}-\mu)]=\gamma_k$,

γ_k 表示 Y_t 与滞后 k 期的 Y_{t-k}的协方差(或自协方差),即 Y_t 的均值、方差和协方差都与时间 t 无关(γ_k 虽然不一定是零,但它只与两个时期的间隔 k 有关,而与具体的时间 t 无关),则说 Y_t 是**弱平稳**或**协方差平稳**的,也简单地说 Y_t 是平稳的.如果 Y_t 不是平稳的,为了化 Y_t 为平稳,还须区分两种情形:

(1) 如果 Y_t 有某种上升或下降趋势,而这一趋势可由一个时间的函数 $f(t)$,比方说,$f(t)=\alpha+\beta t$ 来表示,从 Y_t 中减去 $f(t)$,有

$$Y_t-f(t)=Y_t-(\alpha+\beta t)=u_t,$$

或

$$Y_t = \alpha + \beta t + u_t, \tag{5-1}$$

就能得到一个平稳序列 u_t,则称 Y_t 为**趋势平稳过程**(TSP).因此,对于一个 TSP,在做回归分析时,加进一个时间因素作为解释变量即可避免由于不平稳性引起的分析问题.

(2) 如果 Y_t 是按如下机制生成的,即

$$Y_t = Y_{t-1} + u_t, \quad 即 \quad \Delta Y_t = Y_t - Y_{t-1} = u_t, \tag{5-2}$$

或

$$Y_t = \beta + Y_{t-1} + u_t, \quad 即 \quad \Delta Y_t = Y_t - Y_{t-1} = \beta + u_t. \tag{5-3}$$

其中,$u_t \overset{\text{i.i.d.}}{\sim} D(0,\sigma^2)$是所谓"**白噪声**"(white noise)平稳序列,表示 u_t 是按某个以 0 为均值、σ^2 为方差的分布 D 独立而同分布的;β 是所谓"漂移"(drift)常数,顾名思义,则可称 Y_t 为**差分平稳过程**(DSP).这是一个一阶差分平稳过程,又称随机游动(random walk)模型,其中(5-3)式代表一种带漂移的随机游动(random walk with a drift),而(5-2)式则代表一种无漂移的随机游动.显然,当 Y_t 是 DSP 时,取它的差分 ΔY_t 便得到一个平稳过程.

从(5-3)式容易导出一个酷似(5-1)式的形式

$$\begin{aligned} Y_t &= \beta + Y_{t-1} + u_t \\ &= \beta + (\beta + Y_{t-2} + u_{t-1}) + u_t \\ &= \cdots = Y_0 + \beta t + \sum_{i=1}^{t} u_i, \end{aligned}$$

但(5-3)式和(5-1)式的根本区别在于误差项的方差.对 TSP 来说是 $E(u_t^2) = \sigma^2$,而对 DSP 来说则是 $E\left[(\sum u_i)^2\right] = t\sigma^2$.后者将随 t 的增大而趋于无穷大,所以说 TSP 是一均值不平稳过程,而 DSP 是一方差不平稳过程.检验一个时间序列 Y_t 是否平稳,区分 TSP 和 DSP 是重要的.当 Y_t 是一种带漂移的 DSP 时,它的变化趋势是随时改变的,就是说,它的趋势是随机地变化的,并不像 TSP 那样有一个稳定的长期变化趋势.因此,当没有把握分清楚是 TSP 或是 DSP 时,尤其不能单凭对时间 t 的回归而轻易地做过长的长期预测.当过程是 DSP 时,对误差项的一个冲击(比如从 u_t 变到 $u_t + c$)将产生永久性的长期效果(tc).

5.2 单位根检验

为了检验时间序列 Y_t 是否平稳,并区分它是 TSP 还是 DSP,可通过回归方程

$$Y_t = \alpha + \beta t + \rho Y_{t-1} + u_t \tag{5-4}$$

的估计来检验假设 $H_0: \rho = 1$ 且 $\beta = 0$,及其相对的假设 $H_1: |\rho| < 1$(或 $\rho < 1$).适当的检验方法是本篇 1.2.4 所讨论的 F 检验.但若将(5-4)式等同于一个一阶自相关模型

$$Y_t = \delta_0 + \delta_1 t + u_t, \quad u_t = \rho u_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{i.i.d.}}{\sim} D(0,\sigma^2),$$

或将其合并为

$$Y_t = [\delta_0(1-\rho) + \rho\,\delta_1] + \delta_1(1-\rho)t + \rho Y_{t-1} + \epsilon_t,$$

则有

$$\alpha=\delta_0(1-\rho)+\rho\delta_1 \quad 和 \quad \beta=\delta_1(1-\rho) \quad ((5\text{-}4)式中的 u_t 视同 \varepsilon_t.)$$

此时 $\rho=1$ 意味着 $\beta=0$,从而可利用 t 统计量仅检验 $H_0:\rho=1$. 若接受 H_0,则认为序列 Y_t 是 DSP[参见(5-3)式];若拒绝 H_0,则认为它是 TSP[参见(1-7)式].

按照博克斯-詹金斯(Box-Jenkins)时间序列分析的一套方法论,如果序列 Y_t 本身不是平稳的,则总可以通过取差分(一阶或高阶差分)的方法得到一个平稳序列. Y_t 是否平稳,关键在于判别模型

$$Y_t=\rho Y_{t-1}+u_t \tag{5-5}$$

中的 $\rho=1$ 或 $|\rho|<1$ 这两个关系式中哪一个适用. 如果是 $|\rho|<1$,就表示序列 Y_t 是平稳的;如果 $\rho=1$ 适用,就再检验其一阶差分

$$\Delta Y_t=\rho\Delta Y_{t-1}+\Delta u_t$$

中的 $\rho=1$ 或 $|\rho|<1$,直至得到一个适当高阶的差分,比方说 d 阶差分 $\Delta^d y_t$,是平稳序列为止. 这里

$$\Delta^d y_t=\Delta(\Delta^{d-1}y_t),\cdots,\Delta y_t=y_t-y_{t-1}.$$

如果从 d 开始,$\Delta^d Y_t$ 为平稳序列,而 $\Delta^k y_t(k<d)$ 皆为不平稳序列,则称 Y_t 为 d 阶积整序列,记为 $I(d)$. 根据经验,宏观经济变量如国民收入、货币供给量、股市指数等时间序列大多为一阶积整序列(即 $I(1)$),少数为零阶(即 $I(0)$)和二阶(即 $I(2)$)或高阶积整序列. 检验一个序列是几阶积整的方法,就是检验该序列及其各阶差分的自回归方程中的 ρ 系数是属于"$\rho=1$"还是"$|\rho|<1$",这种检验叫做**单位根检验(unit root tests)**.

常用的 t 检验(或 F 检验)并不适合于用来检验类似方程(5-5)或(5-4)中的 ρ 是否等于 1. 当 $\rho=1$ 时,ρ 的最小二乘估计并不围绕 1 而是围绕比 1 要小的数来分布(虽然这个负的偏误会随样本的增大而减小). Dickey 和 Fuller 曾用蒙特卡罗(Monte Carlo)法表算出用以检验对立于 $|\rho|<1$ 的假设 $H:\rho=1$ 的 t 统计量临界值. 这些临界值一般要比标准的 t 的临界值大许多. 为了区别于标准的 t 检验,人们称这种有关单位根的 t 检验为 **D-F 的 t 检验**或 τ 检验.

单位根检验虽可直接针对诸如(5-4)式或(5-5)式中的 ρ 而检验,但通常更多地采用以下的变换形式:

$$Y_t=\alpha+\rho Y_{t-1}+u_t,\quad H:\rho=1,$$

从方程两边减去 Y_{t-1} 得

$$\Delta Y_t=\alpha+(\rho-1)Y_{t-1}+u_t=\alpha+\delta Y_{t-1}+u_t,\ H:\delta=0. \tag{5-6}$$

类似地,对 $Y_t=\alpha+\beta t+\rho Y_{t-1}+u_t,H:\rho=1$ 也可变为

$$\Delta Y_t=\alpha+\beta t+\delta Y_{t-1}+u_t,\quad H:\delta=0. \tag{5-7}$$

一般地说,在不同的回归方程(5-6)式和(5-7)式中,δ 的 LS 估计 $\hat{\delta}$ 是不相同的. 如果仅考虑不带漂移的随机游动:

$$Y_t=\rho Y_{t-1}+u_t,\quad H:\rho=1$$

或

$$\Delta Y_t = \delta Y_{t-1} + u_t, \quad H:\delta = 0, \tag{5-8}$$

则 δ 的 LS 估计 $\hat{\delta}$ 又将是不同的.因此,检验 $H:\delta = 0$ 的 t 统计量临界值将随回归方程的不同设定而有所区分.一般的单位根 t 检验临界值表都区分三种情形:

(1) 回归方程中仅含滞后项,如(5-8)式.

(2) 兼含常数项和滞后项,如(5-6)式.

(3) 兼含常数项、滞后项和趋势项,如(5-7)式.查表时,应加注意.参看表 5-1.

表 5-1[①] 单位根 t 检验临界值

样本大小	$AR(1)$		$AR(1)$带常数		$AR(1)$带常数与趋势	
	1%	5%	1%	5%	1%	5%
25	−2.66	−1.95	−3.75	−3.00	−4.38	−3.60
50	−2.62	−1.95	−3.58	−2.93	−4.15	−3.50
100	−2.60	−1.95	−3.51	−2.89	−4.04	−3.45
250	−2.58	−1.95	−3.46	−2.88	−3.99	−3.43
500	−2.58	−1.95	−3.44	−2.87	−3.98	−3.42
∞	−2.58	−1.95	−3.43	−2.86	−3.96	−3.41

①本表取自 Maddala G.S. Introduction to Econometrics.1992.第 606 页

考虑到 u_t 可能有自相关(一阶或高阶)的情形,可在检验单位根的回归方程中适当加进一些差分的滞后项,如对带常数和趋势的方程(5-7),把它扩充为

$$\Delta Y_t = \alpha + \beta t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \varepsilon_t,$$

$$\varepsilon_t \sim D(0,\sigma^2),$$

其中 m 根据经验决定.这时相应的对假设 $\delta = 0$ 的 D-F 检验称为增广的(augmented)D-F 检验,简写 ADF 检验.但 ADF 检验和D-F检验,不管是 t 或是 F 检验,都可利用渐近相同的临界值.

5.3 谬误回归与协积回归

考虑长度相同的时间序列.一般地说,d 阶积整序列与 $k(k \leqslant d)$阶积整序列的线性组合是 d 阶积整序列.例如 1 阶积整序列与 1 阶积整序列的线性组合是 1 阶积整序列.由此可知,不平稳序列与平稳(零阶积整)或不平稳序列的线性组合是不平稳序列.鉴于时间序列的线性回归就是时间序列之间的一个线性组合,因此,涉及不平稳时间序列的回归,一般地说,将给出一个不平稳的误差序列.这样便不适宜于对回归计算的结果做标准的 t 或 F 检验了.类似此情形,如果错误地应用 t 或 F 检验,把不显著的回归计算结果当做显著,就会造成谬误回归(spurious regression).

然而,有一种可能,当一个 d 阶积整过程 $I(d)$对另一 $I(d)$做回归(特别是 d = 1)时,得到的误差序列却是 $I(0)$即平稳过程.这时称这两个 d 阶积整过程是"协积

(或协整)"的(cointegrated),或者说它们之间有协积关系.通过一定的线性组合,两个不平稳序列的非平稳性互相抵消了,好比两条基本上平行的曲线相减之后,变成了一条水平线.

有协积关系的两个(或多个)时间序列所做的回归(指两个或多个时间序列的一个线性组合给出一个平稳误差序列)称为**协积回归**(cointegrating regression).协积回归方程中的系数(常数和回归系数)构成一个向量,称为**协积向量**(cointegrating vector).

5.3.1 协积检验的一般步骤

假定要做时间序列 Y_t 对时间序列 X_t 的线性回归,则首先要通过单位根检验以确定 Y_t 和 X_t 各是多少阶积整序列(或过程).设 Y_t 是 $I(d)$, X_t 是 $I(d')$,如果 $d = d' = 0$,则表明它们都是平稳的.这时做 Y_t 对 X_t 的回归就不涉及平稳性的问题.如果 $d = d' \geq 1$,则应做**协积检验**以决定回归误差序列是否平稳,即是否为 $I(0)$.假定回归模型为

$$Y_t = \alpha + \beta X_t + u_t,$$

因误差 u_t 是不可观测的,用 LS 回归残差 $\widehat{u}_t$ 代替 u_t 做 $\widehat{u}_t$ 的单位根检验时,临界值还要作适当调整.通常采用两种方法:

(1) 计算

$$\text{CRDW} = \sum(\widehat{u}_t - \widehat{u}_{t-1})^2 \Big/ \sum \widehat{u}_t^2,$$

但不按 D-W 的 d 值表查相对于 $H: d = 2$ 的临界值,而是按 CRDW(Cointegrating Regression D-W)的缩写的 d 值表查相对于 $H: d = 0$ 的临界值(参见参考文献[7]和[2]).

(2) 按照 Engle 和 Granger 给出的临界值表,查 t 比率的临界值,对 $\Delta\widehat{u}_t = \delta\widehat{u}_{t-1} + \epsilon_t$ 中的 δ 检验 $H: \delta = 0$①.

当 $d \neq d'$ 时, Y_t 和 X_t 不会有协积关系.这时应分别将 Y_t 和 X_t 化为平稳序列或化为同阶积整后再做回归.

以上讲的是做协积检验的正式步骤.Granger 和 Newbold 曾提出一个经验规则:如果最小二乘回归的结果是 $R^2 > d$,就可以怀疑回归的谬误性.这时有必要对相关的时间序列做协积检验.

5.3.2 协积概念的某些应用

(1) 对不平稳序列 Y_t 的合理预期 Y_t^* 应符合以下两个要求:

1) Y_t 与 Y_t^* 不但有协积关系,而且协积系数等于 1;

2) 误差 $Y_t - Y_t^*$ 是白噪声过程,因此,可通过回归方程

---

① 参见 Engle R, Granger C. Econometrica. 1987, 55, 251 ~ 276

$$Y_t = \beta_0 + \beta_1 Y_t^* + \varepsilon_t$$

检验假设 $H:\beta_0=0,\beta_1=1$. 然而,如果 Y_t 和 Y_t^* 都是 $I(d)$, $d\neq 0$,则还要在 H 被接受的情形下进一步检验 $Y_t - Y_t^* = \varepsilon_t$ 是否是 $I(0)$. 即使 ε_t 是 $I(0)$,也还不保证 ε_t 没有序列相关,因此还要做 ε_t 是否有序列相关的检验.

(2) 金融市场有效性应排除从一种价格推测另一种价格的可能性. 例如,不可能从黄金价格的变化推测白银价格的变化;不可能从对一种货币的外汇率推测对另一种货币的外汇率. 因此,黄金价格和白银价格应无协积关系,不同货币的外汇率也无协积关系,等等. 但另一方面,远期汇率在理论上可作为将来的现期汇率的预测元(regressor),它们之间应有协积关系.

5.4　协积与误差纠正机制

按照博克斯-詹金斯方法,任何不平稳的时间序列都可通过取差分(一阶或高阶差分)而化为平稳序列. 对平稳序列做回归分析,就不会出现上述谬误回归的问题. 例如, Y_t 和 X_t 都是一阶积整即 $I(1)$序列时,取它们的一阶差分,再做回归,就可按通常的方法进行统计推断. 然而,这样做虽解决了统计方法问题,却带来了经济分析上的缺陷. 因为差分只反映经济变量的短期变化,而经济学家更关心的却是经济变量之间的长期均衡关系. 经济理论总是考虑一个或多个经济变量是怎样达到均衡或最优状态的. 这只能由 Y_t 本身(或者说 Y_t 的水平值)对 X_t 本身(X_t 的水平值)的回归表现出来. 也就是说,差分与差分之间的回归代替不了水平值与水平值之间的回归. 为了研究变量之间的长期均衡关系,协积关系仍然是重要的. 当 Y_t 和 X_t 有协积关系时,相应的协积回归残差 $\widehat{u}_t$ 是一平稳序列. 这样, ΔY_t, ΔX_t 和 $\widehat{u}_t$ 都是平稳序列,便可做如下的所谓**误差纠正模型**(error-correcting model)的回归分析:

$$\Delta Y_t = \alpha_0 + \alpha_1 \Delta X_t + \alpha_2 \widehat{u}_{t-1} + \varepsilon_t. \tag{5-9}$$

其中

$$\widehat{u}_t = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_t. \tag{5-10}$$

(5-10)式中的 $\widehat{\beta}$和 $\widehat{u}$分别是 Y_t 对 X_t 协积回归中的系数和残差(代表一种均衡误差)的 LS 估计. 当 Y_t 的变化和 X_t 的变化达到均衡时,误差 u_t 应为零. 因此, $\widehat{u}_t$ 代表一种失衡(disequilibrium)度量. 在(5-9)式中, $\widehat{u}_{t-1}$的系数 α_2 就代表时期 t 对前一时期 $t-1$ 的失衡的一种纠正力度,称为误差纠正系数. 比如说, $\alpha_2 = -0.1$,就表示本期 Y_t 的改变值 ΔY_t,除了随 X_t 的改变量 ΔX_t 而改变的部分外,还包含了纠正前一期失衡部分 $\widehat{u}_{t-1}$的 10%. 这样, Y_t 的短期变化 ΔY_t,既联系到 X_t 的短期变化 ΔX_t,也反映了对偏离长期趋势的纠正作用. 但是,还应强调指出,误差纠正模型的理论依据是以 Y_t 和 X_t 有协积关系作为前提的.

误差纠正模型早在 20 世纪 60 年代即被应用,只是到了 70 年代末 80 年代初才得到"协积"理论上的支持.

5.5 向量自回归方法

在建立一个或一组结构性回归方程时必须事先区分外生(自)变量和内生(应)变量.这种区分虽说是经过一定的经济理论分析的,但毕竟带有主观性.为了避免这种主观性,可把所有变量一律当做内生变量来处理,让每一变量都取决于它自己的过去值和其他有关变量的过去值,从而在时间上有一个从过去到现在的先后导向,使之成为一个**向量自回归**(vector auto-regression,简记 VAR)**模型**.例如,考虑由国民总产值 GNP 和货币供给量 M 两个变量构成的向量,相应的向量自回归形式如:

$$\mathrm{GNP}_t = \alpha + \sum_{i=1}^{m}\beta_i \mathrm{GNP}_{t-i} + \sum_{i=1}^{m}\gamma_i M_{t-i} + u_{1t},$$

$$M_t = \omega + \sum_{i=1}^{m}\theta_i \mathrm{GNP}_{t-i} + \sum_{i=1}^{m}\lambda_i M_{t-i} + u_{2t},$$

其中误差项 u_t 在时间序列特别是 VAR 文献中常常称为冲击量(impulse)或新生量(innovation).VAR 和结构性联立方程组相比,可以说是一种非理论性(a-theoretic)模型;它无须对变量作任何先验性约束以保证模型的可识别性.因每个方程的右端都是前定变量,假定 u_t 不是自相关序列,则可直接用 LS 法估计模型.为了预测的目的,仍要求相应的 u_t 是平稳序列.

VAR 模型在应用上的困难主要是滞后项的个数 m 难于决定.如 m 取得较大,则多重共线性的影响将使个别的参数估计误差大幅度增加,致使一方面个别参数的 t 检验均不显著;而另一方面,对一组参数或对整个回归方程的 F 检验则可能异常显著.有鉴于此,在分析一个 VAR 模型时,往往不去问一个变量的变化对另一个变量的影响如何,而是考虑当一个误差项发生变化,或者说模型受到某种冲击时,将对各个变量产生一些什么影响.这种分析方法称为**脉冲响应函数**(impulse response function,简记 IRF)**法**.

和一维的回归方程一样,VAR 也要考虑序列的平稳性问题和序列之间是否存在协积关系的问题.例如,如果 GNP_t 和 M_t 有协积关系,就可考虑它们相互的误差纠正机制(ECM).为说明简单起见,不妨取 $m=1$.设 GNP_t 和 M_t 有协积关系

$$\mathrm{GNP}_t - gM_t = v_t, \quad v_t \sim D(0,\sigma^2),$$

可考虑如下的 ECM:

$$\Delta\mathrm{GNP}_t = \rho_1 v_{t-1} + \mathrm{lagged}(\Delta\mathrm{GNP}_t, \Delta M_t) + \varepsilon_{1t},$$

$$\Delta M_t = \rho_2 v_{t-1} + \mathrm{lagged}(\Delta\mathrm{GNP}_t, \Delta M_t) + \varepsilon_{2t}.$$

其中 lagged(·,·)表示 $\Delta\mathrm{GNP}_t$ 和 ΔM_t 的滞后项的线性函数.当然,在此 ECM 中,还可放进一些滞后更多期的失衡误差项(如 v_{t-2}, v_{t-3},等等).

当向量的维数大于 2 时,可能有不止一个协积关系,而且,若干个协积关系的线性组合仍是一个协积关系,那么,在应用中如何选取适当的协积关系进行预测,就要看哪一个协积关系有较好的经济含义并且最能说明问题.

对于维数较大的情形,协积回归的估算检验和选择是比较复杂的.详见参考文

献[1].

参考文献

1 Charamza W W , Deadman D F. New directions in econometric practice. Vermont: Edward Elgar, 1992.

2 Engle R F , Granger C W J. Cointegration and error correction: representation, estimation and testing. Econometrica, 1987, 55(2): 251 ~ 276

3 Greene W H. Econometric analysis. 3rd ed., New Jersey: Prentice Hall, 1997.

4 Johnson N , Kotz S. Distributions in statistics univariate distributions. New York: Wiley, 1970.

5 Maddala G S. Introduction to econometrics. 2nd ed. New York: Macmillan, 1992.

6 Maddala G S. Limited-dependent and qualitative variables in econometrics. New York: Cambridge University Press, 1983.

7 Mills T W. Time series techniques for economists. New York: Cambridge University Press, 1990.

8 Sargan J D , Bhargava A. Testing residuals from least squares, regression for being Generated by the gaussian random walk. Econometrica, 1983, 51(1): 153 ~ 147

9 林少宫,李楚霖.简明经济统计与计量经济.上海:上海人民出版社,1993.

10 胡代光,高鸿业主编.现代西方经济学辞典.北京:中国社会科学出版社,1996, 305 ~ 344

·经济数学卷·

第2篇

数理经济

编　者　张金水
审校者　张顺明

目　录

引　言

数理经济学一般被定义为包括数学概念和方法在经济学特别是在经济理论中的各种应用.它还可以被定义为采用更多的数学方法来描述的经济学.

数理经济学的研究方法可简要概括为列方程与解方程.第1步:列方程,也就是用数学公式来描述经济系统中的基本环节,如定义了效用函数、产品需求函数、生产函数、供给函数、要素需求函数、消费函数、储蓄函数等;进一步就是用联立方程组来描述经济系统中各变量间的因果关系.第2步:解方程并讨论解的5个基本问题,即解的存在性、稳定性、合理性、能控性、一定时间内到达合理轨道的能达性.

数理经济学的开创性工作是库洛特(A. Cournot)1838年完成的.瓦尔拉斯(M.E.L. Walras)列出了产品市场供求一般均衡的联立方程组.阿罗(K.J. Arrow)、德布鲁(G. Debreu)等证明了产品市场一般均衡解的存在性与唯一性。其后斯卡夫(H.E.Scarf)等给出了求市场均衡点的具体算法.冯·诺伊曼(J.von Neumann)、列昂惕夫(W.W.Leontief)等创建了线性多部门模型来进行产品市场一般均衡分析,等等.数理经济学的理论与实践目前正处在迅速发展之中.在深度上正将有限种产品的有限维商品空间推广到无穷维商品空间,在广度上将仅包含产品市场的一般均衡分析推广到包含产品市场、资本市场、劳动市场、货币市场、国际贸易市场等的一般均衡分析,并讨论相应的平衡增长与最优增长问题.

1　消费者理论

1.1　效用函数

经济学就是研究如何利用有限资源合理安排生产,生产出来的产品在消费者中如何进行合理分配,以达到人类现在和将来的最大满足.人类的最大满足就是经济系统的目标.数理经济学的首要任务就是给出人类最大满足的数学表达式.由于人类的满足与所有个人的满足有关,因此经济系统目标值 $U_{总}$ 应该是个人满意度 $U_1,\cdots,U_m$ 的函数,即

$$U_{总}=f(U_1,\cdots,U_m),\tag{1-1}$$

其中,U_i 是第 i 个人的满意度。

第 i 个人的满意度或个人幸福函数与许多因素有关.一个人的幸福与他享受到的物质量有关,与闲暇、健康、安全感、荣誉感、知足感、婚姻与家庭、妒忌心等有关.因此要给出个人幸福函数圆满的数学表达式是一件极为困难的事.由于在影响一个人生活水平的众多因素中,最主要的是享受到的物质量与闲暇的多少,因此在

构造个人幸福函数表达式时应抓住主要矛盾.如果用U表示某个人的满意度,x_i表示享受到的第i种消费品的数量,T表示所享受到的闲暇时间,那么便有

$$U = U(x_1,\cdots,x_n,T), \tag{1-2}$$

其中的U称为效用函数,它可以理解为狭义的个人幸福函数.如果不考虑闲暇时间T,那么(1-2)式变为

$$U = U(x_1,\cdots,x_n). \tag{1-3}$$

1. 商品空间

设有n种商品,第i种商品量为x_i,它为非负实数,即$x_i \in \mathbf{R}_+$,n种商品量$x = [x_1,\cdots,x_n]^{\mathrm{T}}$的值取自$n$维商品空间$\mathbf{R}_+^n$,即$x \in \mathbf{R}_+^n$,其中$\mathbf{R}_+^n$为$n$维实空间$\mathbf{R}^n$中正象限.

2. 偏好关系

n维商品空间中任取两点x与y,它们表示两组商品,消费者可以比较其优劣.记$x \succ y$表示x比y好,$x \succcurlyeq y$表示x比y好或一样好,$x \sim y$表示x与y一样好或称之为无差异.

3. 用效用函数反应偏好关系

设$x \in \mathbf{R}_+^n$,$y \in \mathbf{R}_+^n$,可以依消费者偏好来构造出相应的效用函数:对满意度大的商品组合赋以较大的效用函数值,满意度一样的两个商品组合赋以相同的效用函数值.即$x \preccurlyeq y$等价于$U(x) \leqslant U(y)$,$x \prec y$等价于$U(x) < U(y)$,$x \sim y$等价于$U(x) = U(y)$.

当消费者拥有商品x时,所得到的满意度大小数值很难确定,在实际应用中人们也不关心它的具体数值.因此,只要函数$U(x)$能反映消费者的偏好顺序,就可以将它当作该消费者的效用函数.

例1　当某个消费者拥有的两种消费品量分别为x_1与x_2时,如下函数反映相同的偏好顺序:$U(x_1,x_2) = x_1^{0.3}x_2^{0.6}$,$U(x_1,x_2) = x_1^{0.1}x_2^{0.2}$,$U(x_1,x_2) = 0.3\ln x_1 + 0.6\ln x_2$,$U(x_1,x_2) = 100x_1^{0.3}x_2^{0.6}$,$U(x_1,x_2) = Ax_1^a x_2^b$(其中,$A,a,b$为正实数,$a:b = 1:2$).如果其中某一个函数反映该消费者对消费品的偏好顺序,则其余几个也反映相同的偏好顺序.若某一个函数反映该消费者的偏好顺序,则就可以将其当作他的效用函数.因此以上几个函数都可以看作消费者的效用函数.

从上例可以看出,若用$U = Ax_1^a x_2^b$来近似消费者对消费品的满意度,则参数A,a,b的大小反映了满意度的大小,a与b的比值反映了偏好顺序.在实际应用中,A,a,b的具体数值测不出来,也不必去关心它的大小.但a与b的比值却可以通过消费者行为的实际数据来测定.

4. 效用函数的性质

效用函数反映消费者对商品的偏好顺序与选择行为.依消费者的消费行为,一般假设效用函数$U(x)$是严格凹函数,即设$x,y \in \mathbf{R}_+^n$,当$x \neq y$,$0 < \alpha < 1$时,成立

$$U(\alpha x + (1-\alpha)y) > \alpha U(x) + (1-\alpha)U(y). \tag{1-4}$$

$U(x)$是凹函数意味着它的二阶偏导数的Hessian阵是负定的.即

$$\frac{\partial^2 U(\boldsymbol{x})}{\partial \boldsymbol{x}^2} = \begin{bmatrix} \frac{\partial^2 U(\boldsymbol{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 U(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 U(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 U(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix} \tag{1-5}$$

是负定的.负定对称阵主对角线上元素小于零,即

$$\frac{\partial^2 U(\boldsymbol{x})}{\partial x_i^2} < 0, \quad i = 1, \cdots, n. \tag{1-6}$$

上式称为效用函数边际递减法则.

5. 效用函数的数学表达式

一般很难给出效用函数具体的数学表达式来准确反映人们的消费行为.但在实际应用中,为了计算上的方便,往往用各种形式的数学公式来近似描述消费者的消费行为.

常用的效用函数表达式有如下几种:

第 1 种　对数线性型

$$U(\boldsymbol{x}) = A(x_1 - a_1)^{\beta_1} \cdots (x_n - a_n)^{\beta_n}. \tag{1-7}$$

其中,$a_i \geqslant 0$ 为第 i 种消费品的最低需要量;A,β_i 为正常数,且 $\beta_1 + \cdots + \beta_n < 1$.

第 2 种　CES(constant elasticity of substitution)型

$$U(\boldsymbol{x}) = \left(\sum_{i=1}^{n} a_i^{1/\sigma} x_i^{(\sigma-1)/\sigma} \right)^{\delta/(\sigma-1)}. \tag{1-8}$$

其中,a_i、σ,$\delta(\delta < \sigma)$ 为正常数.

与上式类似的效用函数有

$$U(\boldsymbol{x}) = \left(\sum_{i=1}^{n} a_i^{1/\sigma} (x_i - b_i)^{(\sigma-1)/\sigma} \right)^{\delta/(\sigma-1)}. \tag{1-9}$$

1.2 需求函数

若消费者消费支出总额为 M,那么在市场价格 $p_1, \cdots, p_n$ 之下购买各种消费品的数量由如下数学模型求解:

$$\begin{cases} \max U(\boldsymbol{x}), \\ \text{s.t.} \quad p_1 x_1 + \cdots + p_n x_n = M, \\ \qquad\quad x_i \geqslant 0, \quad i = 1, \cdots, n. \end{cases} \tag{1-10}$$

上式取极值的必要条件又称为**效用最大法则**.即

$$\begin{cases} \text{当} \frac{\partial U(\boldsymbol{x})}{\partial x_i} < \lambda p_i \text{时}, \quad x_i = 0; \\ \text{当} \frac{\partial U(\boldsymbol{x})}{\partial x_i} = \lambda p_i \text{时}, \quad x_i \geqslant 0; \\ p_1 x_1 + \cdots + p_n x_n = M. \end{cases} \tag{1-11}$$

如果将(1-10)式中约束条件 $x_i \geqslant 0$ 去掉,那么(1-11)式中可相应去掉第1式及第2式中的 $x_i \geqslant 0$.

若给出效用函数 $U(x)$ 的具体数学表达式,通过求解(1-11)式可得到相应的**消费品需求函数**

$$x_i = x_i(p_1, \cdots, p_n, M), \quad i = 1, \cdots, n. \tag{1-12}$$

例2 如果效用函数如(1-7)式所示,那么需求函数为

$$p_i x_i = p_i a_i + \frac{\beta_i}{\beta_1 + \cdots + \beta_n}(M - p_1 a_1 - \cdots - p_n a_n), \quad i = 1, \cdots, n. \tag{1-13}$$

上式所示的需求函数称为**线性支出系统**,即第 i 种消费品支出额 $p_i x_i$ 是价格及收入 M 的线性函数.

例3 如果效用函数如(1-8)式所示,那么相应需求函数为

$$x_i = \frac{a_i p_i^{-\sigma} M}{\sum_{i=1}^{n} a_i p_i^{1-\sigma}}, \quad i = 1, \cdots, n. \tag{1-14}$$

例4 如果某消费者在1年中拥有闲暇时间为 H,那么工作时间为 $(1-H)$ 人年,当每人年工资率为 w 时该消费者消费行为由下式求解:

$$\begin{cases} \max\ U(x_1, \cdots, x_n, H) = A(x_1 - a_1)^{\beta_1} \cdots (x_n - a_n)^{\beta_n}(H - b)^h, \\ \text{s.t.} \quad p_1 x_1 + \cdots + p_n x_n = w(1 - H), \end{cases} \tag{1-15}$$

其中,A, a_i, β_i, h 为正常数;b 为闲暇最低需要量.

从(1-15)式求出对第 i 种消费品最优需求量 x_i 为

$$p_i x_i = p_i a_i + \frac{\beta_i}{\beta_1 + \cdots + \beta_n + h}(w - p_1 a_1 - \cdots - p_n a_n - wb), \tag{1-16}$$

对闲暇 H 的需求量为

$$H = b + \frac{h}{\beta_1 + \cdots + \beta_n + h}\left(1 - b - \frac{p_1 a_1 + \cdots + p_n a_n}{w}\right). \tag{1-17}$$

该消费者工作时间为 $L = (1 - H)$ 人年,在市场机制下它是产品价格 p_i 与劳动力价格 w 的函数,称之为**劳动供给函数**.

1.3 间接效用函数

从(1-3)式可知,效用是享受物质量 x_i 的函数,而在市场机制下享受的物质量又是价格与收入 M 的函数,因此效用可间接看做是价格与收入 M 的函数.

$$V(p_1, \cdots, p_n, M) = U[x_1(p_1, \cdots, p_n, M), \cdots, x_n(p_1, \cdots, p_n, M)], \tag{1-18}$$

上式中 $V(p_1, \cdots, p_n, M)$ 称为**间接效用函数**,它有如下几个特点:

(1) 齐次性 价格及消费总支出 M 同时扩大或缩小 k 倍,满意度或效用值不变.即

$$V(kp_1, \cdots, kp_n, kM) = V(p_1, \cdots, p_n, M). \tag{1-19}$$

(2) 当其他条件不变时,消费总支出 M 上升(或下降),满意度 V 也随之增加

(或下降).

(3) 当其他条件不变时,任一种产品价格上升(或下降),V值将下降(或上升).

一旦给出符合上述几个条件的间接效用函数,消费者行为就可由下式描述:

$$\begin{cases}\text{求极值} & V(p_1,\cdots,p_n,M),\\ \text{约束} & p_1x_1+\cdots+p_nx_n=M.\end{cases} \tag{1-20}$$

上式极值的必要条件为

$$x_i=\frac{-\partial V/\partial p_i}{\partial V/\partial M},\quad i=1,\cdots,n. \tag{1-21}$$

给出一种 V 的具体表达式,便可确定相应的需求函数表达式.

例 5 间接加对数系统.

令间接效用函数为

$$V(p_1,p_2,M)=a_1\left(\frac{M}{p_1}\right)^{b_1}+a_2\left(\frac{M}{p_2}\right)^{b_2}, \tag{1-22}$$

其中,$a_ib_i>0,b_i>-1,i=1,2$.

将(1-22)式代入(1-21)式,可求得需求函数

$$x_i=\frac{a_ib_i(M/p_i)^{b_i+1}}{a_1b_1(M/p_1)^{b_1}+a_2b_2(M/p_2)^{b_2}}.$$

上式所示的需求函数称为**间接加对数系统**.

1.4 需求比较静态分析

当给出严格凹的效用函数 $U(\boldsymbol{x})$,再利用效用最大法则或在预算约束下求效用最大的极值必要条件,所求出的需求函数 $x_i=x_i(p_1,\cdots,p_n,M)$ 应满足如下几个基本条件:

1. 零度齐次性

当价格与收入 M 同时扩大或缩小 k 倍时,需求量不变,即

$$x_i(kp_1,\cdots,kp_n,kM)=x_i(p_1,\cdots,p_n,M).$$

2. 恩格尔条件

$$\eta_1\sigma_1+\cdots+\eta_n\sigma_n=1,$$

其中,$\eta_i=\dfrac{\partial x_i}{\partial M}\times\dfrac{M}{x_i}$ 为需求收入弹性,$\sigma_i=\dfrac{p_ix_i}{M}$ 为第 i 种消费品支出占总支出比例.

3. 古诺条件

$$-\sigma_j=\xi_1\sigma_1+\cdots+\xi_n\sigma_n,\quad j=1,\cdots,n,$$

其中,$\xi_i=\dfrac{\partial x_i}{\partial p_j}\times\dfrac{p_j}{x_i}$ 为第 i 种产品需求量对第 j 种产品价格的交叉弹性.

4. 加总条件

$$p_1x_1+\cdots+p_nx_n=M.$$

加总条件即预算约束.由于在极值必要条件中包含预算约束的方程,因此所求出的需求函数必然满足加总条件.

例6 线性需求函数

$$x_i = a_{1i} p_1 + \cdots + a_{ni} p_n + b_i M, \quad i = 1, \cdots, n, \tag{1-23}$$

不满足齐次条件.

例7 对数线性需求函数

$$x_i = A_i p_1^{a_{1i}} p_2^{a_{2i}} \cdots p_n^{a_{ni}} M^{b_i}, \quad i = 1, \cdots, n, \tag{1-24}$$

当 $a_{1i} + a_{2i} + \cdots + a_{ni} + b_i = 0$ 时,满足零度齐次性条件,但不满足加总条件.

在实际应用中,人们往往凭直觉给出(1-23)式、(1-24)式所示的需求函数.当应用历史数据估计其中参数时,虽然有时能通过统计学检验,但这些需求函数结构与经济学基本假设不符合.因此,当价格与收入作较大范围变动时,应用这些需求函数所做的预测将与实际情况严重不符.

2 生产者理论

2.1 生产函数

在工厂和农村,人们每日每时都在不断生产各种产品,要生产这些产品必须投入各种生产资料或生产要素.它们主要包括:① 劳动者:具有不同文化层次、不同年龄的工人、农民、管理者、知识分子与科研人员等.② 劳动工具:机器、厂房、设备等各种固定资产.③ 各种原材料.④ 土地、矿山、森林等各种资源.可把生产过程中投入的物资或人统称为要素,投入的各要素与产出产品之间因果关系的数学表达式称为**生产函数**.

如果一种生产过程仅生产一种产品,则称该生产过程为**无联合生产**.如果一种生产过程同时生产多种产品,则称该生产过程为**有联合生产**.

无联合生产的生产函数一般可用下式表示:

$$Y = f(z_1, \cdots, z_n, K_1, \cdots, K_m, L_1, \cdots, L_s), \tag{2-1}$$

其中,$z_1, \cdots, z_n$ 表示各种原材料等的中间投入数量;$K_1, \cdots, K_m$ 表示投入的各种固定资本与资源的数量;$L_1, \cdots, L_s$ 表示投入的各种不同层次的劳动工时的数量.

要想给出(2-1)式的具体数学表达式,来准确反映生产过程中投入量与产出量之间的数量关系,是一件十分困难的事情.在实际应用中,人们往往用各种类型的数学公式来近似反映现实的生产过程.

常用的生产函数具体数学表达式有如下几种:

1. 柯布-道格拉斯(Cobb-Douglas)型生产函数

$$Y = AK^a L^b, \tag{2-2}$$

其中,K 为固定资本投入量,L 为劳动工时投入量,A, a, b 为正常数.当 $a + b = 1$ 时,称之为规模报酬不变的生产函数;当 $a + b > 1$ 时,称之为规模报酬递增的生产函数;当 $a + b < 1$ 时,称之为规模报酬递减的生产函数.

2. 列昂惕夫(Leontief)型生产函数

$$Y = \min\left(\frac{z_1}{a_1}, \cdots, \frac{z_n}{a_n}, \frac{K_1}{b_1}, \cdots, \frac{K_m}{b_m}, \frac{L_1}{l_1}, \cdots, \frac{L_s}{l_s}\right). \tag{2-3}$$

上式表明,投入的各种要素应成恰当比例.当投入的各要素量分别为

$$z_i = a_i y, \quad i = 1, \cdots, n;$$
$$K_j = b_j y, \quad j = 1, \cdots, m;$$
$$L_k = l_k y, \quad k = 1, \cdots, s$$

时,产出 $Y = y$.

3. 柯布-道格拉斯型与列昂惕夫型相互嵌套的生产函数

$$\begin{cases} Y = \min\left(\frac{z_1}{a_1}, \cdots, \frac{z_n}{a_n}, V\right), \\ V = AK_1^{b_1} \cdots K_m^{b_m} L_1^{l_1} \cdots L_s^{l_s}, \end{cases} \tag{2-4}$$

上式中,投入的原材料要成恰当比例,而投入的固定资本、资源和劳动力相互之间可以替代.比如,生产一块糖,需要一张包糖纸与 2g 糖的中间投入成恰当比例,但生产糖可以采用机械化、半机械化、手工劳动为主等各种方式,即可以用机器来代替人的手工劳动.

4. CES 型生产函数

$$Y = A(aK^{\sigma} + bL^{\sigma})^{\delta/\sigma}, \tag{2-5}$$

其中,K 为固定资本投入量,L 为劳动工时投入量,(2-5) 式中设只有 1 种固定资本与 1 种劳动力;A, a, b, δ 等为给定正常数,$-\infty < \sigma < 1$.当 $\delta = 1$ 时,(2-5) 式为规模报酬不变的生产函数,$\delta > 1$ 与 $\delta < 1$ 时分别为规模报酬递增与规模报酬递减的生产函数.

5. 列昂惕夫型与 CES 型相互嵌套的生产函数

$$\begin{cases} Y = \min\left(\frac{z_1}{a_1}, \cdots, \frac{z_n}{a_n}, V\right), \\ V = A(b_1 K_1^{\sigma} + \cdots + b_m K_m^{\sigma} + l_1 L_1^{\sigma} + \cdots + l_s L_s^{\sigma})^{1/\sigma}, \end{cases} \tag{2-6}$$

上式中,投入的原材料 $z_1, \cdots, z_n$ 要成恰当比例,而投入的固定资本、资源、劳动力则可以互相替代.

当(2-6) 式中 $\sigma = -\infty$ 时,(2-6) 式化为(2-3) 式所示的列昂惕夫型的生产函数.当(2-6) 式中 $\sigma = 0$ 时,(2-6) 式化为(2-4) 式所示的柯布-道格拉斯型与列昂惕夫型相互嵌套的生产函数.可以给出(2-1) 式所示生产函数的各种各样的具体形式,在实际应用时应根据实际情况选择适当的类型.

有联合生产的生产函数一般可用下式表示:

$$f(y_1, \cdots, y_n, z_1, \cdots, z_n, K_1, \cdots, K_m, L_1, \cdots, L_s) = 0, \tag{2-7}$$

其中,$y_1, \cdots, y_n$ 为 n 种产品产出量,其余变量含义与(2-1) 式相同.

(2-7) 式可以有各种各样的具体形式,例如,有联合生产的列昂惕夫型生产函数为

$$(y_1,\cdots,y_n)-(g_1,\cdots,g_n)\times\min\left(\frac{z_1}{a_1},\cdots,\frac{z_n}{a_n},\frac{K_1}{b_1},\cdots,\frac{K_m}{b_m},\frac{L_1}{l_1},\cdots,\frac{L_s}{l_s}\right)=0,\tag{2-8}$$

上式表明,当投入要素成恰当比例,

$$z_i = ka_i,\quad i=1,\cdots,n;$$
$$K_j = kb_j,\quad j=1,\cdots,m;$$
$$L_q = kl_q,\quad q=1,\cdots,s$$

时,产出为 $y_i = kg_i, i=1,\cdots,n$.其中 g_i 为给定常数.

2.2　供给函数与要素需求函数

考虑(2-1)式所示的生产函数,当产出为 Y 时,收入为 pY,p 为产品价格,为产出 Y 需投入的原材料、固定资产、资源、劳动工时等所付出的成本

$$C = p_1z_1+\cdots+p_nz_n+r(p_1K_1+\cdots+p_nK_n+ p_{n+1}K_{n+1}+\cdots+p_mK_m)+w_1L_1+\cdots+w_sL_s,\tag{2-9}$$

其中,$p_i, i=1,\cdots,n$,为 n 种产品的价格;$p_j, j=n+1,\cdots,m$,为资源的价格;r 为市场资本统一利润率;$w_q, q=1,\cdots,s$,为各种劳动力的工资率.

生产者的利润 Π 为收入减去成本,即

$$\Pi = pY - C,\tag{2-10}$$

在完全竞争的市场机制下,生产者谋求利润最大应符合如下极值必要条件:

$$\begin{cases} p\dfrac{\partial Y}{\partial z_i} = p_i, & i=1,\cdots,n; \\ p\dfrac{\partial Y}{\partial K_j} = rp_j, & j=1,\cdots,m; \\ p\dfrac{\partial Y}{\partial L_q} = w_q, & q=1,\cdots,s. \end{cases}\tag{2-11}$$

其中,rp_j 为资本或资源的租金.

(2-11)式所示的极值必要条件又称为生产者利润最大法则.

用求解(2-11)式的极值必要条件,可以求出生产过程的最优投入量,它们都是市场价格、资本或资源的租金、工资率等的函数.即

$$\begin{cases} z_i = z_i(p_1,\cdots,p_m,rp_1,\cdots,rp_m,w_1,\cdots,w_s); \\ K_j = K_j(p_1,\cdots,p_m,rp_1,\cdots,rp_m,w_1,\cdots,w_s); \\ L_q = L_q(p_1,\cdots,p_m,rp_1,\cdots,rp_m,w_1,\cdots,w_s). \end{cases}\tag{2-12}$$

上式称为**要素需求函数**.将(2-12)式代入生产函数(2-1)式中,可以看出产出量 Y 也是市场价格、资本或资源的租金、工资率等的函数.即

$$Y = Y(p_1,\cdots,p_m,rp_1,\cdots,rp_m,w_1,\cdots,w_s),\tag{2-13}$$

上式称为**产品供给函数**.

例1　设(2-1)式的生产函数为如下柯布-道格拉斯型:

$$Y = Az_1^{a_1}\cdots z_n^{a_n}K_1^{b_1}\cdots K_m^{b_m}L_1^{l_1}\cdots L_s^{l_s}, \tag{2-14}$$

其中,中间投入 z_i 的价格为p_i,固定资本及资源 K_j 的价格为rp_j,劳动工时 L_q 的价格为 w_q,产品 Y 的价格为p.

当生产函数具有规模报酬递减的性质,即

$$e = a_1 + \cdots + a_n + b_1 + \cdots + b_m + l_1 + \cdots + l_s < 1 \tag{2-15}$$

时,产品供给函数为

$$Y = \left[A\left(\frac{a_1}{p_1}\right)^{a_1}\cdots\left(\frac{a_n}{p_n}\right)^{a_n}\left(\frac{K_1}{rp_1}\right)^{b_1}\cdots\left(\frac{K_m}{rp_m}\right)^{b_m}\left(\frac{L_1}{w_1}\right)^{l_1}\cdots\left(\frac{L_s}{w_s}\right)^{l_s}p^e\right]^{1/(1-e)}, \tag{2-16}$$

其中,e 如(2-15)式中所定义.

投入各要素的需求函数为

$$\begin{cases} z_i = p\left(\dfrac{a_i}{p_i}\right)Y, & i = 1,\cdots,n; \\ K_j = p\left(\dfrac{b_j}{rp_j}\right)Y, & j = 1,\cdots,m; \\ L_q = p\left(\dfrac{l_q}{w_q}\right)Y, & q = 1,\cdots,s. \end{cases} \tag{2-17}$$

其中,Y 如(2-16)式所示,它是产品价格、固定资本或资源的租金、工资率的函数,因此投入各要素也是产品价格、资源的租金、工资率的函数.

当生产函数具有规模报酬不变的性质时,$e = 1$,e 如(2-15)式所示,产品价格 p 与要素价格之间的关系为

$$p = \frac{1}{A}\left(\frac{p_1}{a_1}\right)^{a_1}\cdots\left(\frac{p_n}{a_n}\right)^{a_n}\left(\frac{rp_1}{b_1}\right)^{b_1}\cdots\left(\frac{rp_m}{b_m}\right)^{b_m}\left(\frac{w_1}{l_1}\right)^{l_1}\cdots\left(\frac{w_s}{l_s}\right)^{l_s}. \tag{2-18}$$

上式即为产品供给函数.但应当注意到,当生产函数具有规模报酬不变的性质时,供给量 Y 不能表示产品价格与要素价格的连续函数.这时供给函数可以表述为:在给定各要素价格时,按(2-18)式计算出来的产品价格记为 p^*.当产品实际价格小于 p^* 时,产品供给量 $Y = 0$;当产品实际价格等于 p^* 时,产品供给量无论多大都可以.

由于产出 Y 难以用一个数学公式表示出它与产品价格、资源的租金、工资率之间的关系,也难以用简洁的数学公式表示出要素需求量与价格、租金、工资率之间的关系,因此在生产函数具有规模报酬不变的性质时,一般仅列出如下的单位产出要素需求函数:

$$\begin{cases} \dfrac{z_i}{Y} = \dfrac{a_i p}{p_i}, & i = 1,\cdots,n; \\ \dfrac{K_j}{Y} = \dfrac{b_j p}{rp_j}, & j = 1,\cdots,m; \\ \dfrac{L_q}{Y} = \dfrac{l_q p}{w_q}, & q = 1,\cdots,s. \end{cases} \tag{2-19}$$

其中,z_i/Y 是产出为1单位时,第 i 种中间投入品的需求量;K_j/Y 是产出为1单位时

第 j 种固定资本或资源的需求量；L_q/Y 为产出为1单位时第 q 种劳动力需求量.

例2　设(2-1)式的生产函数为如下的CES型：

$$Y = A[a_1 z_1^{(\sigma-1)/\sigma} + \cdots + a_n z_n^{(\sigma-1)/\sigma} + b_1 K_1^{(\sigma-1)/\sigma} + \cdots + b_m K_m^{(\sigma-1)/\sigma} + l_1 L_1^{(\sigma-1)/\sigma} + \cdots + l_s L_s^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)}, \tag{2-20}$$

其中各变量含义与(2-14)式同.

由于(2-20)式是具有规模报酬不变的生产函数，因此，单位产出要素需求函数为

$$\begin{cases} \dfrac{z_i}{Y} = A^{\sigma-1}\left(\dfrac{a_i p}{p_i}\right)^{\sigma}, & i = 1,\cdots,n; \\ \dfrac{K_j}{Y} = A^{\sigma-1}\left(\dfrac{b_j p}{r p_j}\right)^{\sigma}, & j = 1,\cdots,m; \\ \dfrac{L_q}{Y} = A^{\sigma-1}\left(\dfrac{l_q p}{w_q}\right)^{\sigma}, & q = 1,\cdots,s. \end{cases} \tag{2-21}$$

产品价格 p 与要素价格之间的关系(或称之为供给函数)为

$$p = A^{-1}\left[a_1\left(\frac{a_1}{p_1}\right)^{\sigma-1} + \cdots + a_n\left(\frac{a_n}{p_n}\right)^{\sigma-1} + b_1\left(\frac{b_1}{r p_1}\right)^{\sigma-1} + \cdots + b_m\left(\frac{b_m}{r p_m}\right)^{\sigma-1} + l_1\left(\frac{l_1}{w_1}\right)^{\sigma-1} + \cdots + l_s\left(\frac{l_s}{w_s}\right)^{\sigma-1}\right]^{-1/(\sigma-1)}. \tag{2-22}$$

当(2-21)式与(2-22)式中的 $\sigma = 1$ 时，它们分别化为(2-17)式与(2-18)式的形式，这时生产函数(2-20)式化为(2-14)式的形式. 因此，柯布-道格拉斯型的生产函数仅是CES型生产函数的特殊情况.

例3　如果生产函数具有(2-3)式的形式，那么单位产出要素需求函数为

$$\begin{cases} z_i/Y = a_i, & i = 1,\cdots,n; \\ K_j/Y = b_j, & j = 1,\cdots,m; \\ L_q/Y = l_q, & q = 1,\cdots,s. \end{cases} \tag{2-23}$$

此时，产品价格 p 与要素价格之间的关系(或称之为供给函数)为

$$p = a_1 p_1 + \cdots + a_n p_n + r(b_1 p_1 + \cdots + b_m p_m) + w_1 l_1 + \cdots + w_s l_s. \tag{2-24}$$

2.3　供给比较静态分析

供给比较静态分析主要讨论供给函数及要素需求函数的基本性质，即研究市场价格的变化将怎样影响产品供给量与要素投入量的变化.

在消费者理论中，给出严格凹的效用函数，在消费者预算约束之下依效用最大法则可得到相应的消费品需求函数. 在生产者理论中，给出生产函数，在产品与要素价格确定之后依生产者利润最大法则可得到相应的产品供给函数与要素需求函数. 因此，消费者理论与生产者理论在方法上有许多类似之处. 例如，当生产函数为规模报酬递减的严格凹函数时，所得到的产品供给函数、要素需求函数与消费品需

求函数有许多类似的性质.基本的性质有如下几条:

1° 当生产函数为增函数且是严格凹函数,即边际产出递减时,在其他条件不变的情况下,产品价格上升(或下降)将引起该种产品供给量上升(或下降),即 $\partial Y/\partial p > 0$.这个性质可以从(2-16)式的具体例子中得以证实.

2° 条件同 1°.一定存在某些或全部要素,当它的价格上升时将使供给量下降.例如,由(2-16)式可知

$$\begin{cases}\partial Y/\partial p_i < 0, & i = 1,\cdots,n;\\ \partial Y/\partial(rp_j) < 0, & j = 1,\cdots,m;\\ \partial Y/\partial w_q < 0, & q = 1,\cdots,s.\end{cases} \tag{2-25}$$

3° 条件同 1°.某种要素价格上升将引起该种要素需求量下降.例如,从(2-16)式与(2-17)式可知

$$\begin{cases}\partial z_i/\partial p_i < 0, & i = 1,\cdots,n;\\ \partial K_j/\partial(rp_j) < 0, & j = 1,\cdots,m;\\ \partial L_q/\partial w_q < 0, & q = 1,\cdots,s.\end{cases} \tag{2-26}$$

4° 条件同 1°.产品供给函数及要素需求函数满足零度齐次条件,即当产品与要素价格同时扩大或缩小相同倍数时,产品供给量与要素需求量不变.

在实际应用中,往往采用规模报酬不变的生产函数.从(2-18)式、(2-19)式可以看出产品供给函数及要素需求函数有如下基本性质:

1° 从(2-18)式可知,当所有要素价格同时扩大 α 倍时,产品价格也将扩大相同倍数.这意味着供给函数有齐次性.即当产品与要素价格同时扩大 α 倍时,对产品供给量没有影响.

2° 从(2-18)式可知,任意一种要素价格上升将引起产品价格上升.

3° 从(2-19)式可知,任意一种要素价格上升(或下降)将引起该种要素单位产出要素需求量下降(或上升).

4° 从(2-19)式可知,产品与要素同时扩大 α 倍,将不影响单位产出要素需求量的变化.

3　一般均衡理论

3.1　瓦尔拉斯一般均衡理论

供给与需求的平衡是经济学中要讨论的核心内容.一般地说,一种产品的供给量与需求量与其他各种产品价格都有关系.现设有 n 种产品,第 i 种产品总需求量 D_i 是 n 种产品价格的函数:$D_i = D_i(p_1,\cdots,p_n)$,$i = 1,\cdots,n$.第 i 种产品总供给量也是 n 种产品价格的函数:$S_i = S_i(p_1,\cdots,p_n)$,$i = 1,\cdots,n$.供求平衡时各产品价格与供求量由如下联立方程求解:

$$D_i(p_1,\cdots,p_n) = S_i(p_1,\cdots,p_n), \quad i = 1,\cdots,n. \tag{3-1}$$

在经济学中,瓦尔拉斯最早给出了列写产品市场一般均衡联立方程组的方法,从而奠定了一般均衡理论的基础.

下面介绍瓦尔拉斯一般均衡模型.

1. 消费者效用最大法则与需求函数

设有 m 个消费者,n 种产品,则第 i 个人效用函数为

$$U_i = U_i(x_{i1},\cdots,x_{ij},\cdots,x_{in}),$$

其中,x_{ij} 表示第 i 个消费者享受到的第 j 种消费品的数量.

第 i 个消费者对 n 种产品的初期占有量分别为 $x_{i1}^*,\cdots,x_{in}^*$. 在市场价格下,他要出售这些产品并购入他喜欢的产品. 产品交换应满足如下的预算约束:

$$p_1x_{i1} + \cdots + p_nx_{in} = p_1x_{i1}^* + \cdots + p_nx_{in}^*. \tag{3-2}$$

消费者在预算约束之下求效用最大,应满足如下的效用最大法则:

$$\frac{\partial U_i}{\partial x_{i1}}\bigg/p_1 = \frac{\partial U_i}{\partial x_{i2}}\bigg/p_2 = \cdots = \frac{\partial U_i}{\partial x_{in}}\bigg/p_n, \tag{3-3}$$

求解(3-2)式与(3-3)式可求出第 i 个消费者对 n 种产品的需求量或需求函数.

2. 生产者利润最大法则与供给函数及要素需求函数

设有 s 个生产者,第 k 个生产者投入 n 种产品中的某些产品来产出 n 种产品中另一些产品. 其生产函数可表示为

$$f_k(\tilde{x}_{k1},\cdots,\tilde{x}_{kj},\cdots,\tilde{x}_{kn}) = 0, \quad k = 1,\cdots,s, \tag{3-4}$$

其中,$\tilde{x}_{kj}$ 表示第 k 个生产者所投入或产出的第 j 种产品的数量.

第 k 个生产者在生产过程中所获得的利润 Π_k 为

$$\Pi_k = p_1\tilde{x}_{k1} + p_2\tilde{x}_{k2} + \cdots + p_n\tilde{x}_{kn},$$

在(3-4)式生产函数约束之下求利润 Π_k 最大的极值必要条件(或称之为利润最大法则)为

$$\frac{\partial f_k}{\partial \tilde{x}_{k1}}\bigg/p_1 = \frac{\partial f_k}{\partial \tilde{x}_{k2}}\bigg/p_2 = \cdots = \frac{\partial f_k}{\partial \tilde{x}_{kn}}\bigg/p_n. \tag{3-5}$$

解(3-4)式与(3-5)式可得到第 k 个生产者产品供给函数与要素需求函数.

3. 市场供求平衡

第 j 种产品在生产过程中投入或产出的总量为 $\tilde{x}_{1j} + \cdots + \tilde{x}_{ij} + \cdots + \tilde{x}_{sj}$,其中,$\tilde{x}_{ij}$ 为第 i 个生产者所投入或产出的第 j 种产品的数量. 对消费者来说,第 q 个消费者购买的第 j 种产品的数量为 x_{qj},其初期占有量为 x_{qj}^*,因此他对第 j 种产品的总需求为 $x_{qj} - x_{qj}^*$,全体 m 个消费者对第 j 种产品总需求应等于生产者总供给,因此有如下平衡方程式:

$$\sum_{i=1}^{s}\tilde{x}_{ij} = \sum_{q=1}^{m}(x_{qj} - x_{qj}^*). \tag{3-6}$$

(3-2) ~ (3-6) 式构成瓦尔拉斯一般均衡模型.现在来分析方程组个数与变量的个数.

对(3-2)式来讲,由于有 m 个消费者,故相应有 m 个方程.对(3-3)式来讲,每个消费者有 $n-1$ 个方程,故相应有 $m\times(n-1)$ 个方程.对(3-4)式来讲,s 个生产者相应有 s 个方程.对(3-5)式来讲,相应有 $s\times(n-1)$ 个方程.对(3-6)式来讲,n 种产品相应有 n 个平衡方程.因此方程总个数为

$$m+m\times(n-1)+s+s\times(n-1)+n=m\times n+s\times n+n;$$

相应地,x_{ij} 有 $m\times n$ 个变量,$\tilde{x}_{kj}$ 有 $s\times n$ 个变量,价格 p_j 有 n 个变量,因此变量总数正好也是 $m\times n+s\times n+n$ 个.在瓦尔拉斯一般均衡模型中,方程的个数与变量的个数相等.

关于瓦尔拉斯一般均衡模型应注意如下几个要点:

(1) 方程个数与未知数个数相等时解的存在性与唯一性问题.一般地讲,瓦尔拉斯一般均衡模型中只有 $m\times n+s\times n+n-1$ 个方程相互独立,因此只能求出产品间相互比价 $p_1:p_2:\cdots:p_n$.相互间比价是唯一的.

(2) 瓦尔拉斯一般均衡模型只考虑产品市场的供求平衡问题,没有涉及资本品市场、货币市场、国际贸易市场等,没有考虑税收收入与财政支出问题,也没有考虑环境污染与保护问题以及教育与科研问题等,全面考虑这些问题属于现代应用一般均衡理论.

现代应用一般均衡理论目前仍在迅速发展之中.

3.2 存在一般均衡解的瓦尔拉斯条件与一般均衡解的存在性

关于产品市场供求平衡联立方程组的列写原则与解存在的条件要点如下:

1° 设有 n 种商品,这些商品可以是消费品,也可以是生产过程中投入的生产要素,如原料、资本品、劳动工时、土地、各种其他资源,等等.各种商品价格分别为 $p_1,p_2,\cdots,p_n$.

2° 消费者在预算约束之下求效用最大可得到消费品需求函数,生产者在生产函数约束之下求利润最大可得到要素需求函数与产品供给函数.消费品需求函数与要素需求函数相加可得到产品总需求函数,它是价格 $p_1,\cdots,p_n$ 的函数.设第 i 种商品总需求量为 $D_i=D_i(p_1,\cdots,p_n)$,总供给量为 $S_i=S_i(p_1,\cdots,p_n)$,则可得到供求平衡的 n 个方程为

$$D_i(p_1,\cdots,p_n)=S_i(p_1,\cdots,p_n),\quad i=1,\cdots,n. \tag{3-7}$$

3° 需求函数与供给函数满足齐次条件,即当所有价格都扩大或缩小同一倍数时,需求与供给量不变:

$$D_i(kp_1,\cdots,kp_n)=D_i(p_1,\cdots,p_n),$$
$$S_i(kp_1,\cdots,kp_n)=S_i(p_1,\cdots,p_n).$$

4° 如果消费者效用函数为严格凹函数,那么消费品需求函数便为连续函数.

5° 如果生产函数为规模报酬递减的严格凹函数，那么产品供给函数与要素需求函数皆为连续函数.

应当指出，在实际应用中往往采用规模报酬不变的生产函数，它不再是严格凹函数，这时产品供给函数与要素需求函数将不是连续函数，这种情况下一般均衡解的存在性证明将困难得多.

6° 方程

$$\sum_{i=1}^{n} p_i[D_i(p_1,\cdots,p_n) - S_i(p_1,\cdots,p_n)] = 0 \tag{3-8}$$

称为瓦尔拉斯条件.它的含义是:各种收入应等于各种支出之和.

7° 考虑(3-7)式所示市场供求平衡方程，如果供求函数满足连续性、齐次性、瓦尔拉斯条件，那么(3-7)式有解存在，即市场存在供求平衡点.

8° (3-7)式一般是相关的，即只能求出各产品间相互比价 $p_1:p_2:\cdots:p_n$.

下面举例说明上述要点.

例1 某封闭社会共有10 000人，只进行两种生产活动:织布与种田.有两个生产者分别经营这两种生产活动.织布生产活动的生产函数为

$$y_1 = \sqrt{L_1}, \tag{3-9}$$

其中，L_1 为投入的劳动工时量，y_1 为布的产出量.

种田生产活动的生产函数为

$$y_2 = 2\sqrt{L_2}, \tag{3-10}$$

其中，L_2 为投入劳动工时量，y_2 为粮食产出量.(注:本例的目的仅在于说明一般均衡方程的列写方法及平衡点存在的条件，因此对生产函数作了简化.如果该封闭社会有许多生产活动，每种生产活动都要投入许多种要素，那么将涉及许多变量.更复杂与更切合实际的模型在可计算一般均衡模型中讨论).

设每个人偏好函数都一样，为

$$U_i = x_{i1}^{1/4}x_{i2}^{1/4}(H_i - 0.4)^{1/4}, \tag{3-11}$$

其中，x_{i1} 为第 i 个人享受到布的数量;x_{i2} 为第 i 个人享受到粮食的数量;H_i 为第 i 个人拥有闲暇时间的数量;0.4 人年为拥有闲暇时间的最低需要量.

产品拿到市场上进行交换，问:当市场供求平衡时，布的价格 p_1、粮食价格 p_2 及工资率 w 的比值 $p_1:p_2:w=?$

当布的生产者投入为 L_1 时，成本为 wL_1，其中 w 为工资率;产出为 y_1，收入为 p_1y_1.生产者的目标是获得利润 Π_1 极大化，即

$$\max \Pi_1 = p_1y_1 - wL_1. \tag{3-12}$$

通过求解(3-9)式与(3-12)式利润最大必要条件，可得到布的供给函数为

$$y_1 = p_1/(2w), \tag{3-13}$$

以及投入的要素 L_1 的需求函数

$$L_1 = p_1^2/(4w^2). \tag{3-14}$$

当粮食生产者投入 L_2 劳动工时时，成本为 wL_2，产出为 y_2，相应收入为 p_2y_2，生产者要获得极大化利润 Π_2，即

$$\max \Pi_2 = p_2 y_2 - wL_2. \tag{3-15}$$

通过求解(3-10)式与(3-15)式的利润最大必要条件，可得到粮食的供给函数为

$$y_2 = 2p_2/w, \tag{3-16}$$

以及投入的要素 L_2 的需求函数

$$L_2 = p_2^2/w^2. \tag{3-17}$$

生产者获得的总利润

$$\begin{aligned} \Pi_1 + \Pi_2 &= p_1 y_1 + p_2 y_2 - wL_1 - wL_2 \\ &= \frac{p_1^2}{2w} + \frac{2p_2^2}{w} - \frac{p_1^2}{4w} - \frac{p_2^2}{w} = \frac{p_1^2}{4w} + \frac{p_2^2}{w}. \end{aligned} \tag{3-18}$$

瓦尔拉斯条件说明：各种收入应等于各种支出之和. 在本例中，销售收入为 $p_1 y_1 + p_2 y_2$，工资收入为 wL^s（L^s 为劳动总可供量 10 000 人年减去闲暇时间总需求量），生产中的工资支出为 $wL_1 + wL_2$，消费总支出为 $p_1 x_1 + p_2 x_2$（x_1 与 x_2 分别为两种产品消费总需求量），因此瓦尔拉斯条件为

$$p_1 y_1 + p_2 y_2 + wL^s = wL_1 + wL_2 + p_1 x_1 + p_2 x_2, \tag{3-19}$$

上式可改写为

$$p_1(x_1 - y_1) + p_2(x_2 - y_2) + w(L_1 + L_2 - L^s) = 0, \tag{3-20}$$

即为(3-8)式所示的瓦尔拉斯条件.

第 i 个人在预算约束之下达到效用最大的模型为

$$\begin{cases} \max\ U_i = x_{i1}^{1/4} x_{i2}^{1/4} (H_i - 0.4)^{1/4}, \\ \text{s.t. } p_1 x_{i1} + p_2 x_{i2} = w(1 - H_i) + a_i(\Pi_1 + \Pi_2). \end{cases} \tag{3-21}$$

上式中，$L_i^s = 1 - H_i$ 为第 i 个人劳动工时供给量，$L_1^s + \cdots + L_{10000}^s = L^s$；$a_i$ 为转移支付系数. 由于本例中没有公共消费支出、投资支出等，因此利润收入 $\Pi_1 + \Pi_2$ 应转移支付给消费者才能符合瓦尔拉斯条件，故 $a_1 + \cdots + a_{10000} = 1$.

通过求解(3-21)式的效用最大必要条件，可求得第 i 个人两种产品的需求函数及闲暇需求量：

$$\begin{cases} p_1 x_{i1} = 0.2w + \dfrac{a_i}{3}\left(\dfrac{p_1^2}{4w} + \dfrac{p_2^2}{w}\right); \\ p_2 x_{i2} = 0.2w + \dfrac{a_i}{3}\left(\dfrac{p_1^2}{4w} + \dfrac{p_2^2}{w}\right); \\ wH_i = 0.6w + \dfrac{a_i}{3}\left(\dfrac{p_1^2}{4w} + \dfrac{p_2^2}{w}\right). \end{cases} \tag{3-22}$$

全社会相应总需求量为

$$x_1 = \sum_{i=1}^{10000} x_{i1},\ x_2 = \sum_{i=1}^{10000} x_{i2},\ H = \sum_{i=1}^{10000} H_i;$$

劳动力总供给量为

$$L^s = 10000 - H;$$

产品市场供求平衡方程为

$$x_1 = y_1, \quad x_2 = y_2;$$

劳动市场供求平衡方程为

$$L^s = L_1 + L_2.$$

从(3-13)、(3-14)、(3-16)、(3-17)、(3-22)式可求得市场供求平衡方程组

$$\frac{2000w}{p_1} + \frac{1}{3p_1}\left(\frac{p_1^2}{4w} + \frac{p_2^2}{w}\right) = \frac{p_1}{2w}, \tag{3-23}$$

$$\frac{2000w}{p_2} + \frac{1}{3p_2}\left(\frac{p_1^2}{4w} + \frac{p_2^2}{w}\right) = \frac{2p_2}{w}, \tag{3-24}$$

$$4000 - \frac{1}{3w}\left(\frac{p_1^2}{4w} + \frac{p_2^2}{w}\right) = \frac{p_1^2}{4w^2} + \frac{p_2^2}{w^2}, \tag{3-25}$$

不难知道,(3-23) ~ (3-25)式有三个未知数 p_1, p_2, w 及三个方程,且满足齐次条件、连续性条件与瓦尔拉斯条件,此外三个方程只有二个独立,可求出市场比价为

$$p_1 : p_2 : w = 20\sqrt{15} : 10\sqrt{15} : 1.$$

如果在(3-21)式中将预算约束方程改为

$$p_1 x_{i1} + p_2 x_{i2} = w(1 - H_i),$$

那么可以证明将不满足瓦尔拉斯条件,从而不存在平衡点,即市场供求平衡方程组无解.

3.3　阿罗 - 德布鲁一般均衡模型

前面几节给出了列写一般均衡模型的基本思路与方法,其中做了一些假定,如效用函数为严格凹函数且连续可导;生产函数为无联合生产的规模报酬递减的严格凹函数时,相应的供给函数为价格的连续函数,等等.当供给函数与需求函数满足齐次条件、连续性条件及瓦尔拉斯条件时,可以证明均衡点的存在.上述条件有些过于苛刻或与实际尚有较大偏离.例如,在实际应用中一般采用规模报酬不变的生产函数,这时供给函数将不再是连续函数,这将增加证明市场均衡点存在性的难度.在阿罗 - 德布鲁一般均衡模型中,采用集合与映射的语言来描述供求平衡并证明均衡点的存在性.可以说瓦尔拉斯采用的是经典数学知识来描述市场供求平衡,而阿罗 - 德布鲁采用的是近代数学知识来描述它.正由于采用的是近代数学知识,可以对现实经济系统的运动刻画得更加深入,但也更加抽象与缺乏直观性,不利于应用经济工作者对它的了解与掌握.

下面简要介绍阿罗 - 德布鲁一般均衡模型的一些要点.

1. 集值映射

记 $\mathbf{R}^n$ 与 $\mathbf{R}^m$ 分别表示 n 维与 m 维实空间, X 与 Y 分别是它们的子集,即 $x \subset \mathbf{R}^n$ 及 $Y \subset \mathbf{R}^m$,如果对于每一点 $x \in X$,总有一个确定的子集合 $\varphi(x) \subset Y$ 与之对应,则称 φ 为从 X 到 Y 的**集值映射**(或称为点 - 集对应),记作 $\varphi: X \twoheadrightarrow Y$.

初等数学中一元函数 $y = \varphi(x)$,自变量 x 与应变量 y 都是实数,即 $x \in \mathbf{R}, y \in \mathbf{R}$;多元函数 $y = \varphi(x_1, \cdots, x_n) = \varphi(x)$ 中, $x \in \mathbf{R}^n, y \in \mathbf{R}$,它们都是集值映射的特殊情况.

2. 集值映射的连续性

在集值映射中，自变量与应变量都是集合. 直观地说，当自变量集合均匀地变化(不突然变大或缩小)时，应变量集合也作均匀的变化，则称这种集值映射为连续映射. 下面给出集值映射连续性的定义.

定义1 上半连续性 令 $x^1, x^2, \cdots, x^k$ 是收敛于 x^* 的任一序列，其相应映射到集合序列 $\varphi(x^1), \varphi(x^2), \cdots, \varphi(x^k)$. 令 $a^1 \in \varphi(x^1), a^2 \in \varphi(x^2), \cdots, a^k \in \varphi(x^k)$，且 $a^1, a^2, \cdots, a^k$ 收敛于 a，那么，如果 $a \in \varphi(x^*)$，则称 φ 在 x^* 处上半连续. 如果 φ 在 X 处处上半连续，则称 φ 在 X 是上半连续的.

定义2 下半连续性 令 $x^1, x^2, \cdots, x^k$ 是收敛于 x^* 的任一序列，其相应映射到集合序列 $\varphi(x^1), \varphi(x^2), \cdots, \varphi(x^k)$，在 $\varphi(x^*)$ 中任取 $b \in \varphi(x^*)$，总存在序列 $b^k \in \varphi(x^k)$，使得 $b^1, b^2, \cdots, b^k$ 收敛于 b，则称 φ 在 x^* 处下半连续. 如果 φ 在 X 处处下半连续，则称 φ 在 X 是下半连续的.

如果 φ 在 X 既是上半连续，又是下半连续的，则称之为连续的集值映射.

3. 角谷(Kakutani)不动点定理

设 $X \subset \mathbf{R}^n$ 为凸紧集，集值映射 $\varphi: X \rightrightarrows X$ 为从 X 到 X 的集值映射，而且是上半连续的，则 φ 有不动点，即存在 $x^* \in X$，使得 $x^* \in \varphi(x^*)$.

如果 $\varphi: X \to X$ 是 X 到自身的点对点映射，则角谷不动点定理化为布劳威尔(Brouwer)不动点定理.

布劳威尔不动点定理 设 $X \subset \mathbf{R}^n$ 为凸紧集，映射 $\varphi: X \to X$ 是连续的，则 φ 有不动点，即存在 $x^* \in X$，使得 $x^* = \varphi(x^*)$.

不动点定理在证明市场均衡点存在性中起关键作用.

4. 商品空间与偏好关系

设有 n 种商品，由于可供消费的商品数量一般大于或等于零，当用向量 x 表示商品的一个组合时，$x \in \mathbf{R}_+^n$，因此一般称 $\mathbf{R}_+^n$ 为 n 维商品空间. 消费者对商品空间中任二组商品可比较其优劣. 可在 $\mathbf{R}_+^n$ 上定义二元偏好关系 $\preceq$，它满足如下条件：

1° 自反性 对任意 $x \in \mathbf{R}_+^n$，都有 $x \preceq x$.

2° 传递性 当 $x \preceq y, y \preceq z$ 时，有 $x \preceq z$.

3° 完备性 对任意 $x, y \in \mathbf{R}_+^n$，有 $x \preceq y$ 或者 $y \preceq x$.

这里，"$x \preceq y$"的意思是 y 至少同 x 一样好. 如果 $x \preceq y$ 且 $y \preceq x$，则认为 x 与 y 一样好，并记为 $x \sim y$. 如果 $x \preceq y$，但 $x \sim y$ 不成立，这意味着 y 比 x 好，并记为 $x \prec y$.

5. 偏好关系的连续性

定义在 $\mathbf{R}_+^n$ 上的偏好关系称为连续的，如果对任意的 $x \in \mathbf{R}_+^n$，集合 $\{y \in \mathbf{R}_+^n \mid y \preceq x\}$ 与 $\{y \in \mathbf{R}_+^n \mid x \preceq y\}$ 都是 $\mathbf{R}_+^n$ 中的闭集.

6. 偏好关系的单调性与凸性

设 $\preceq$ 是定义在 $\mathbf{R}_+^n$ 上的偏好关系.

(1) 称偏好关系 $\preceq$ 是单调的，如果当 $x \leqslant y$(这表示 y 中各分量 $\geqslant x$ 中相应分量)且 $x \neq y$，则 $y \succ x$.

(2) 称偏好关系 $\preceq$ 是凸的,如果对任意的 $x \in \mathbf{R}_+^n$,集合 $\{y \in \mathbf{R}_+^n \mid x \preceq y\}$ 是凸集.

(3) 称偏好关系 $\preceq$ 是严格凸的,如果它是凸的,且对任意的 $x, y \in \mathbf{R}_+^n, x \neq y$, $x \preceq y$,都有 $x \prec \lambda x + (1-\lambda)y$,其中 $0 < \lambda < 1$.

7. 偏好关系与效用函数

设 $\preceq$ 是定义在 $\mathbf{R}_+^n$ 上的偏好关系,称函数 $u: \mathbf{R}_+^n \to \mathbf{R}_+$ 是表现偏好关系的效用函数,如果 $x \preceq y$,当且仅当 $u(x) \leqslant u(y)$.

显然,$x \sim y$ 当且仅当 $u(x) = u(y)$;而 $x \prec y$,则等价于 $u(x) < u(y)$.

依消费者实际行为,总假设定义在 $\mathbf{R}_+^n$ 上的偏好关系是连续的,可以证明这种情况下存在表现偏好关系 $\preceq$ 的效用函数 $u: \mathbf{R}_+^n \to \mathbf{R}_+$,且 u 是连续函数.

若 $u(x)$ 是表现偏好关系 $\preceq$ 的效用函数,则有如下性质:

1° 偏好关系 $\preceq$ 是单调的,当且仅当 $u(x)$ 是严格递增函数,即当 $x < y$ 时,有 $u(x) < u(y)$.

2° 偏好关系 $\preceq$ 是凸的,当且仅当 $u(x)$ 是拟凹的函数.

3° 偏好关系 $\preceq$ 是严格凸的,当且仅当 $u(x)$ 是严格拟凹的函数.

8. 预算约束集合

在纯交换经济中,消费者在进行交易之前拥有的商品量称为该消费者的初期占有量.设初期占有量为:$a = [a_1, \cdots, a_n]^{\mathrm{T}}$,它为常向量,且 $a \in \mathbf{R}_+^n$. n 种商品价格为 $p = [p_1, \cdots, p_n]$,在市场经济中,消费者出售 a 并购入 x,购入 x 所花费的钱不应超出出售 a 的收入.因此消费者应满足如下预算约束:

$$p \cdot x = p_1 x_1 + \cdots + p_n x_n \leqslant p \cdot a = p_1 a_1 + \cdots + p_n a_n.$$

符合上述约束的点 x 构成预算约束集合 $B(p, a)$.

$$B(p, a) = \{x \in \mathbf{R}_+^n \mid p \cdot x \leqslant p \cdot a\}. \tag{3-26}$$

9. 需求集合与需求映射

消费者在初期占有量为 a 时,在市场价格 p 之下购入 x 使效用最大,可用如下模型描述:

$$\begin{cases} \max & u(x), \\ \text{s.t.} & p \cdot x \leqslant p \cdot a, \quad x \in \mathbf{R}_+^n, \end{cases} \tag{3-27}$$

上式最优解集合 $d(p, a)$ 称为该消费者的需求集合.有时将(3-27)式记为

$$d(p, a) = \arg\max\{u(x) \mid x \in B(p, a)\}, \tag{3-28}$$

需求集合 $d(p, a)$ 与价格 p 有关,即给定一个价格 $p \in \mathbf{R}_+^n$,对应有一个需求集合 $d(p, a) \subset \mathbf{R}_+^n$,这种对应称为 $\mathbf{R}_+^n$ 到自身的集值映射,并称为需求映射 $d: \mathbf{R}_+^n \rightrightarrows \mathbf{R}_+^n$.

10. 纯交换经济的均衡配置

设有 m 个人,记为 $I = \{1, 2, \cdots, m\}$,第 i 个人初期占有量为 $a^i \in \mathbf{R}_+^n$,在市场价格与预算约束之下该消费者购买量 x^i 为需求集合 d^i 中某一点,也就是说购入量 x^i 应使得第 i 个人效用 $u^i(x^i)$ 最大.商品总可供量为 $a = a^1 + a^2 + \cdots + a^m$,总需求量为 $x = x^1 + x^2 + \cdots + x^m$.若所有消费者在市场价格 p^* 之下谋各自效用最大且达供求平衡 $x^* = a$,则称这时的价格 p^* 与需求量 x^* 分别为均衡价格与均衡配

置.

11. 纯交换经济均衡配置的存在性

设 m 个消费者初期占有量总和 $\boldsymbol{a} \in \mathbf{R}_{+}^{n}$ 为严格正的有界常向量,因此可以认为第 i 个消费者消费量限定在商品空间 $\mathbf{R}_{+}^{n}$ 中非空有界凸闭集内,再假定第 i 个消费者售出初期占有 $\boldsymbol{a}^{i}$ 得到收入为 $\boldsymbol{p} \cdot \boldsymbol{a}^{i}$,他将全部收入用来购所需要的商品,即 $\boldsymbol{p} \cdot \boldsymbol{a}^{i} = \boldsymbol{p} \cdot \boldsymbol{x}^{i}$,那么可以证明当第 i 个消费者的效用函数 u^{i} 是连续的且严格拟凹时,$i \in I = \{1, \cdots, m\}$,纯交换经济将存在均衡价格与均衡配置.

经济系统均衡配置存在性的证明在数理经济理论中占据重要地位,阿罗 · 德布鲁等人在 20 世纪 50 年代应用冯 · 诺伊曼等人所创建的对策论及角谷不动点定理等知识给出了产品市场一般均衡配置存在性的严格证明,他们因此分别于 1972 年和 1983 年获得了诺贝尔经济学奖.

纯交换经济中只有消费者,没有生产者.

12. 生产集合

设有 n 种商品,其中有些商品可以作为生产过程中的投入要素,有些则是生产过程中的产出品. 用向量 $\boldsymbol{y} = [y_1, \cdots, y_n]^{\mathrm{T}}$ 表示一个生产活动,若某个分量 $y_i < 0$,表示该生产活动需投入第 i 种商品量 y_i 作为投入要素,若 $y_j > 0$,表示该生产活动产出第 j 种产品量为 y_j,则生产活动 $\boldsymbol{y}$ 的全体构成的集合 T, $T \subset \mathbf{R}^{n}$,称为**生产集合**.

生产集合假设满足如下性质:

1° T 是闭凸集.

2° $\mathbf{R}_{+}^{n} \cap T = \{\mathbf{0}\}$.

3° $(-\mathbf{R}_{+}^{n}) \subset T$.

其中性质 1° 认为生产活动规模可连续变化,且符合规模报酬不变或规模报酬递减的规律;性质 2° 表示 $\boldsymbol{y}$ 中各分量都大于或等于零且至少某些分量大于零是不可能的,这意味着没有投入就不可能有产出;性质 3° 表示允许只有投入却没有产出的情况存在.

13. 利润函数与供给映射

生产活动处于 $\boldsymbol{y}$ 时,所得利润为 $\boldsymbol{p} \cdot \boldsymbol{y} = p_1 y_1 + \cdots + p_n y_n$. 由于生产者总是要采用能获得最大利润的生产活动,因此生产者利润函数 $\Pi(\boldsymbol{p})$ 定义如下:

$$\Pi(\boldsymbol{p}) = \max\{\boldsymbol{p} \cdot \boldsymbol{y} \mid \boldsymbol{y} \in T\}.$$

可获得最大利润的生产活动 $\boldsymbol{y}$ 称为供给量. 供给量 $\boldsymbol{y}$ 与市场价格 $\boldsymbol{p}$ 有关. 供给量 $\boldsymbol{y}$ 所在的集合可用

$$S(\boldsymbol{p}) = \{\boldsymbol{y} \in T \mid \boldsymbol{p} \cdot \boldsymbol{y} = \Pi(\boldsymbol{p})\}$$

表示,或记为

$$S(\boldsymbol{p}) = \arg\max\{\boldsymbol{p} \cdot \boldsymbol{y} \mid \boldsymbol{y} \in T\}.$$

其中 $S(\boldsymbol{p})$ 称为供给映射,它表示每一价格 $\boldsymbol{p}$ 对应一个集合 $S(\boldsymbol{p})$.

14. 考虑消费者与生产者的均衡配置

设全社会有 m 个消费者及 h 个生产者,他们分别组成集合 $I = \{1, \cdots, m\}$ 和集合 $J = \{1, \cdots, h\}$. 第 i 个消费者对 n 种商品初期占有量为 $\boldsymbol{a}^{i}$,效用函数为 u^{i}, $i \in I$.

第 j 个生产者生产集合为 $T^j, j \in J$，利润函数为 $\Pi^j(p)$. 由于没有考虑固定资产投资与经济增长问题，因此所获得的利润应全部用于消费. 设 $\Pi^j(p)$ 转移支付给第 i 个消费者的份额为 $r_{ij}\Pi^j(p)$，那么第 i 个消费者可供消费的总收入为

$$\beta_i(p) = p \cdot a^i + \sum_{j=1}^{h} r_{ij}\Pi^j(p), \quad i \in I,$$

其中，
$$\sum_{i=1}^{m} r_{ij} = 1, \quad j \in J.$$

第 i 个消费者在价格 p 之下购入 x 的消费支出应小于或等于他的总收入 $\beta_i(p)$，他的预算集合可用下式表示：

$$B_i(p) = \{x \in \mathbf{R}_+^n \mid p \cdot x \leqslant \beta_i(p)\}.$$

由于这里没有考虑储蓄的问题，上式中一般可采用等式约束，即 $p \cdot x = \beta_i(p)$.

当所有消费者在市场价格 p^* 之下谋求各自效用最大，生产者谋求各自利润最大，且总需求等于总供给时，总消费需求 x^*、生产总供给 y^* 称为均衡配置，p^* 称为均衡价格. 可用如下式子表示市场均衡配置与均衡价格：

(1) 对于消费者 $i \in I$，其消费需求量 x^i 满足预算约束且达效用最大，即

$$u^i(x^i) = \max\{u^i(z) \mid z \in \beta_i(p^*)\};$$

(2) 对于生产者 $j \in J$，供给量 y^j 满足利润最大法则

$$p \cdot y^j = \Pi^j(p^*) = \max\{p \cdot z \mid z \in T^j\},$$

其中 T^j为第j个生产者生产集合；

(3) 总供求平衡

$$\sum_{i=1}^{m} x^i = \sum_{j=1}^{h} y^j + \sum_{i=1}^{m} a^i.$$

其中，x^i 为第i个消费者消费需求量；y^j 为第j个生产者产品供给量（供给为负时应理解为投入要素需求量）；a^i 为第i个消费者初期占有量.

15. 考虑生产者与消费者时均衡配置的存在性

阿罗与德布鲁等人在20世纪50年代证明了包含消费者与生产者的经济系统均衡配置的存在性. 下面简要说明存在均衡配置消费者与生产者应满足的几个基本条件.

对消费者 $i \in I$ 来讲，其消费量 $x^i \in \mathbf{R}_+^n$，效用函数 u^i 是连续的且严格拟凹的，并且效用函数具有非饱和性，即对任意的 $x^i \in \mathbf{R}_+^n$，存在 $z^i \in \mathbf{R}_+^n$，使得 $u^i(x^i) < u^i(z^i)$.

对生产者 $j \in J$ 来讲，他的生产集合为 $T^j, 0 \in T^j$，并且集合 $V = T^1 + \cdots + T^h$（即 V 为 T^j 中各向量和构成的集合）是闭凸的，允许只投入无产出的情况，即设 $(-\mathbf{R}_+^n) \subset V$. 再假设生产过程是不可逆的，即 $V \cap (-V) = \{0\}$.

若消费者与生产者满足上述条件，此外还满足瓦尔拉斯条件，即消费支出应等于利润收入加消费者出售初期占有的收入，那么，可以证明市场存在供求平衡的均衡配置与均衡价格.

阿罗、德布鲁等人在上述相当宽松的条件下证明了市场均衡配置的存在性问

题,奠定了现代一般均衡理论的基础.但也应当看到上述模型存在许多不足之处.比如:

(1) 仅解决了静态平衡点问题,未涉及均衡增长与最优增长的动态问题;

(2) 未包括税收收入与财政支出,也未考虑货币政策问题;

(3) 未考虑国际贸易问题;

(4) 未考虑教育、科研及环境保护等问题;

(5) 未解决求均衡配置的具体算法,等等.

近几十年来,数理经济学理论与实践处在迅速发展的过程中,上述问题都得到不同程度的解决.数理经济学也成长为一门内容极其广泛与丰富的一门学科.

3.4　列昂惕夫一般均衡模型

考虑(2-3)式所示列昂惕夫型的生产函数,为简便起见,设只有一种劳动力,可将(2-3)式简化为

$$Y = \min\left\{\frac{z_1}{a_1},\cdots,\frac{z_n}{a_n},\frac{K_1}{b_1},\cdots,\frac{K_n}{b_n},\frac{L}{l}\right\}, \tag{3-29}$$

上式是生产一种产品的生产函数.如果有 n 种产品,每种产品由一个生产活动生产出来,那么第 i 种产品的生产函数可表示为

$$Y_i = \min\left\{\frac{z_{1i}}{a_{1i}},\cdots,\frac{z_{ni}}{a_{ni}},\frac{K_{1i}}{b_{1i}},\cdots,\frac{K_{ni}}{b_{ni}},\frac{L_i}{l_i}\right\}, \tag{3-30}$$

当中间投入 z_{ji} 以及固定资本投入 K_{ji} 与劳动工时投入 L_i 成恰当比例,即

$$z_{ji} = a_{ji}y_i,\quad K_{ji} = b_{ji}y_i,\quad L_i = l_iy_i$$

时,产出 $Y_i = y_i$.当第 t 年各种产品产出量为 $\boldsymbol{y}(t) = [y_1(t),\cdots,y_n(t)]^{\mathrm{T}}$ 时,那么第 t 年投入的中间产品为 $\boldsymbol{Ay}(t)$,投入的固定资本量为 $\boldsymbol{By}(t)$,投入的劳动工时量为 $\boldsymbol{ly}(t)$,其中中间投入系数阵 $\boldsymbol{A}$,固定资本投入系数阵 $\boldsymbol{B}$,以及劳动工时投入系数行向量 $\boldsymbol{l}$ 由下面的式子表示:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nn}\end{bmatrix},\quad \boldsymbol{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n}\\ \vdots & & \vdots\\ b_{n1} & \cdots & b_{nn}\end{bmatrix},\quad \boldsymbol{l} = [\,l_1\cdots l_n\,].$$

由于第 t 年投入固定资本量为 $\boldsymbol{By}(t)$,第 $t+1$ 年投入固定资本量为 $\boldsymbol{By}(t+1)$,其增量 $\boldsymbol{B}[\boldsymbol{y}(t+1)-\boldsymbol{y}(t)]$ 即为新增固定资本,这里不考虑固定资本折旧,那么第 t 年产出 $\boldsymbol{y}(t)$ 一方面用于中间投入 $\boldsymbol{Ay}(t)$,另一方面用于固定资本投资 $\boldsymbol{B}[\boldsymbol{y}(t+1)-\boldsymbol{y}(t)]$,余下用于消费.因此有如下的实物平衡方程:

$$\boldsymbol{y}(t) = \boldsymbol{Ay}(t) + \boldsymbol{B}[\boldsymbol{y}(t+1) - \boldsymbol{y}(t)] + \boldsymbol{c}(t). \tag{3-31}$$

上式便是著名的**列昂惕夫动态投入产出模型**.

如果在(3-31)式中不考虑固定资本投资问题,那么可得如下**静态列昂惕夫投入产出模型**:

$$\boldsymbol{y}(t) = \boldsymbol{Ay}(t) + \boldsymbol{c}(t). \tag{3-32}$$

在实际应用中,往往把固定资本投资与折旧放在最终消费 $\boldsymbol{C}(t)$ 中考虑.

现在假设每人年工时消费结构为 $\boldsymbol{d} = [d_1, \cdots, d_n]^{\mathrm{T}}$,那么为产出 $y(t)$ 需投入 $ly(t)$ 人年工时,消费量 $c(t)$ 应为

$$c(t) = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} ly(t) = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} [l_1, \cdots, l_n] y(t) = Ty(t), \tag{3-33}$$

其中,T 为消费系数阵,

$$T = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} [l_1 \cdots l_n] = \begin{bmatrix} d_1 l_1 & \cdots & d_1 l_n \\ \vdots & & \vdots \\ d_n l_1 & \cdots & d_n l_n \end{bmatrix}.$$

将(3-33)式代入(3-31)式,得到动态投入产出模型的另一种表达式为

$$y(t) = Ay(t) + B[y(t+1) - y(t)] + Ty(t). \tag{3-34}$$

由列昂惕夫等人所创建的线性多部门模型近几十年来得到迅速的发展与广泛的应用,例如,(3-31)式所示动态投入产出模型与控制理论新分支——广义动态系统理论相结合,形成广义线性多部门经济系统理论,可方便地讨论动态系统平衡增长、最优增长及运动全过程分析.

4 应用一般均衡理论

4.1 求市场均衡价格与均衡配置的算法

上一章讨论了市场供求平衡的存在性.在实际应用中还要求出均衡价格与均衡配置的具体数值.求解市场均衡价格与均衡配置的算法可以归入应用数学中非线性静态联立方程组求解的范畴.例如,具有 n 种商品的市场供求平衡可由下式描述:

$$\begin{cases} f_1(p_1, \cdots, p_n) = D_1(p_1, \cdots, p_n) - S_1(p_1, \cdots, p_n) = 0; \\ \vdots \\ f_n(p_1, \cdots, p_n) = D_n(p_1, \cdots, p_n) - S_n(p_1, \cdots, p_n) = 0. \end{cases} \tag{4-1}$$

其中,D_i 与 S_i 分别为第 i 种商品需求量与供给量,p_i 为相应价格.

(4-1)式也可简记为

$$f(p) = 0. \tag{4-2}$$

求解(4-2)式非线性方程组,无论是在工程领域还是在经济管理领域都有广泛的应用.例如,在非线性规划中往往包含有类似(4-2)式所示静态方程组的约束条件.因此可以利用非线性规划通用软件(如GAMS软件等)来求解(4-2)式的一般均衡问题.

但是,由于经济系统有其特殊规律,例如(4-1)式中供求函数满足齐次性与瓦尔拉斯条件等,因此经济学家开发出相应软件来求市场均衡解.1960年斯卡夫

(Scarf) 等人给出了在标准单纯形上求解市场均衡点的第一代算法，其后由许多数理经济学家不断完善发展成为第二、三、四…… 代算法，使得求解变量可以更多，速度可以更快. 下面简要介绍在标准单纯形上寻找平衡点算法的基本原理.

1. 标准单纯形

标准单纯形或单位单纯形是如下的集合 S：

$$S = \left\{x \in \mathbf{R}_+^n \,\middle|\, \sum_{i=1}^n x_i = 1, x_i \geqslant 0\right\}, \tag{4-3}$$

其中，x_i 是向量 x 的第 i 个分量.

2. 单纯形

设 $x^1, \cdots, x^r$ 是 n 维实空间 $\mathbf{R}^n$ 中 r 个线性独立的向量，则由此 r 个向量线性组合而生成的向量为

$$x = \alpha_1 x^1 + \cdots + \alpha_r x^r,$$

其中，$\alpha_1 + \cdots + \alpha_r = 1, \alpha_i \geqslant 0, i = 1, \cdots, r$，那么 x 所在的集合 S 叫 $r-1$ 维单纯形.

标准单纯形由 $x^1, \cdots, x^n$ 生成，其中，x^i 的第 i 个分量为 1，其余分量为零.

3. 标准单纯形的端面

标准单纯形的 n 个端面是如下的集合：

$$S_i = \{x \in S \mid x_i = 0\}, i = 1, \cdots, n.$$

4. 单纯形的剖分

将单纯形分为互不重叠的小单纯形 $S^1, \cdots, S^k$，且满足如下两个条件：

(1) $S = S^1 \cup S^2 \cup \cdots \cup S^k$，即单纯形 S 是小单纯形 $S^1, \cdots, S^k$ 的并集；

(2) S^i 与 S^j，$i \neq j$，或不相交，或整个端面相交，或只相交于顶点，则称 $S^1, \cdots, S^k$ 为单纯形 S 的一个剖分.

5. 限制的剖分

设 $K = \{S^1, \cdots, S^m\}$ 是单纯形 S 的一个剖分. 如果除去 S 的顶点之外，胞腔 $S^i (1 \leqslant i \leqslant m)$ 的其余顶点都不在 S 的端面上，则称剖分 K 是一个限制的剖分.

6. 全标号单纯形

设 K 是单纯形 S 的一个限制的剖分，$K = \{S^1, \cdots, S^m\}$，共有 k 个顶点 $V^1, \cdots, V^k (k \geqslant n)$，$S$ 的 n 个顶点标号分别为 $1, \cdots, n$. 其余顶点的标号任选 $1, \cdots, n$ 中之一个，即若 V^j 不是 S 的顶点，则它的标号 $l(V^j) \in \{1, \cdots, n\}$. 如果某个胞腔 S^i 的顶点标号恰好为 $1, \cdots, n$，则称 S^i 为全标号单纯形.

7. 标号定理

设单纯形 S 的一个限制剖分的顶点为 $V^1, \cdots, V^k$，前 n 个顶点是 S 的顶点，赋予标号 $l(V^j) = j, (1 \leqslant j \leqslant n)$，其余顶点的标号 $l(V^i), i > n$，在集合 $\{1, \cdots, n\}$ 中任取，则剖分中必存在一个全标号单纯形.

8. 正常的标号

设 S 是一个标准单纯形或单位单纯形，剖分的顶点为 $V^1, \cdots, V^k$，相应标号为 $l(V^1), \cdots, l(V^k)$，对任一顶点 V^j，若它的标号 $l(V^j) = i$，向量 V^j 的第 i 个分量大于零，则称这种标号规则是正常的.

9. 正常标号下的标号定理

设标准单纯形 S 的剖分 K 的顶点标号规则是正常的，则剖分 K 至少存在一个全标号单纯形.

以上给出了 n 维商品空间 $\mathbf{R}_+^n$ 中标准单纯形剖分的一些性质. 下面利用这些性质来求解(4-1) 式的平衡点. 由于(4-1) 式中供求函数满足齐次条件，若 $p_1,\cdots,p_n$ 是一组解，那么乘以任一正常数 a，则 $ap_1,\cdots,ap_n$ 仍是它的一组解. 因此可以仅在标准单纯形中寻找市场均衡解. 具体方法是将标准单纯形进行剖分，然后给予一定的标号规则，通过寻找全标号单纯形求出市场均衡点.

10. 求市场均衡点的一种算法

第1步：精度选择. 给定正整数 Z，Z 越大意味着剖分胞腔半径越小，从而计算精度越高.

第2步：起始胞腔的选择. 起始胞腔顶点坐标可用如下矩阵表示，每一列表示一个顶点的坐标，凡坐标中有零分量，表示该顶点在标准单纯形端面上.

$$\begin{bmatrix} 0 & 1/Z & 1/Z & \cdots & 1/Z & 1/Z \\ 1/Z & 0 & 1/Z & \cdots & 1/Z & 1/Z \\ 1/Z & 1/Z & 0 & \cdots & 1/Z & 1/Z \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1/Z & 1/Z & 1/Z & \cdots & 0 & 1/Z \\ \dfrac{(Z-n+2)}{Z} & \dfrac{(Z-n+2)}{Z} & \dfrac{(Z-n+2)}{Z} & \cdots & \dfrac{(Z-n+2)}{Z} & \dfrac{(Z-n+1)}{Z} \end{bmatrix}. \tag{4-4}$$

第3步：标准单纯形端面上顶点标号规则的确定. 如果某个胞腔顶点坐标分量中出现有0分量，说明该顶点在标准单纯形端面上. 首先出现0的分量号即为该顶点的标号. 即

$$l(V) = \min\{i \mid V_i = 0, i = 1,\cdots,n\}, \tag{4-5}$$

其中，V 在标准单纯形端面上，V_i 为它的第 i 个分量坐标.

按(4-5) 式标号规则，(4-4) 式中左数第1列所对应的顶点标号为1，第2列所对应的顶点标号为2，…，第 $n-1$ 列所对应的顶点标号为 $(n-1)$. 而第 n 列所对应的顶点在标准单纯形内部，其顶点标号按下面规则进行.

第4步：标准单纯形内部胞腔顶点标号规则的确定. 若顶点 $V = [V_1,\cdots,V_n]^{\mathrm{T}}$ 在标准单纯形的内部，可令 $p_1 = V_1,\cdots,p_n = V_n$，然后分别计算(4-1) 式中的 f_i 值，$i = 1,\cdots,n$. 顶点 V 的标号规则可按下式进行：

$$l(V) = \min\{i \mid f_i > 0, i = 1,\cdots,n\}. \tag{4-6}$$

当胞腔顶点按上述方法标号时，可以证明在标准单纯形内部必存在一个全标号单纯形胞腔.

第5步：胞腔的转移. 在起始胞腔中前 $(n-1)$ 个顶点（对应(4-4) 式的前 $(n-1)$ 列）标号恰好为 $1,\cdots,(n-1)$. 现按第4步确定第 n 个顶点标号，如果标号为 n，则找到一个全标号单纯形胞腔. 但应当指出，由于(4-1) 式经济系统的特殊规律，起始胞腔一般不会是全标号单纯形. 这是因为当 Z 很大时，$p_i = 1/Z, i = 1,\cdots,$

$(n-1)$ 很小,相应地 p_n 则很大. 由于第 n 种产品价格 p_n 相对很大,使得该种产品需求量 D_i 很小,而供给量 S_i 很大,故 $f_i = D_i - S_i < 0$,因此按(4-6)式标号规则,第 n 个顶点标号不会是 n.

如果起始胞腔第 n 个顶点标号与第 j 个顶点(相应(4-4)式的第 j 列,$j \leqslant (n-1)$)标号相同,那么应去掉第 j 列的原有顶点坐标,换上另一个新顶点. 新顶点坐标按下式计算:

$$\begin{cases} b_1^j = b_0^{j+1} + b_0^{j-1} - b_0^j, & \text{当 } 2 \leqslant j \leqslant (n-1) \text{ 时;} \\ b_1^1 = b_0^2 + b_0^n - b_0^1, & \text{当 } j = 1 \text{ 时;} \\ b_1^n = b_0^1 + b_0^{n-1} - b_0^n, & \text{当 } j = n \text{ 时.} \end{cases} \tag{4-7}$$

其中,b_1^j 为新顶点坐标,b_0^{j+1},b_0^{j-1},b_0^j 为原来胞腔顶点坐标.

当用 b_1^j 代替 b_0^j 时,(4-4)式所示矩阵变成另一个矩阵(两矩阵仅第 j 列不同),它相应为新胞腔的 n 个顶点. 这样便从旧胞腔转移到新胞腔. 在新胞腔中再计算 b^j 所对应的顶点的标号,标号规则同(4-5)式或(4-6)式. 如果标号为 n,则找到一个全标号单纯形胞腔;如果标号 $k \leqslant (n-1)$,那么必然有1列标号与之相同,也为 k,如果该列为第 l 列,那么应去掉第 l 列的旧坐标,换上新坐标,新坐标计算方法同(4-7)式(只须将 j 改为 l 即可). 如此不断进行下去,可以证明,由于经济系统自身特性,胞腔转移都在标准单纯形内部进行. 由于胞腔个数有限,转移过程不会循环重复,因此最终将找到一个全标号单纯形胞腔.

第6步:均衡点的计算. 当找到全标号单纯形后,将标号为 i 的顶点记为 a_i. 由于经济系统的特殊性质,全标号单纯形胞腔顶点一般都在标准单纯形内部. 依标号规则有如下性质:

顶点 a_1,标号为1,则 $f_1 > 0$;

顶点 a_2,标号为2,则 $f_1 \leqslant 0, f_2 > 0$;

…

顶点 a_n,标号为 n,则 $f_1 \leqslant 0, \cdots, f_{n-1} \leqslant 0, f_n > 0$.

从以上可知,在 a_1 处 $f_1 > 0$,a_2 处 $f_1 \leqslant 0$,则 a_1 与 a_2 附近必存在一点(依供求函数连续性),使得 $f_1 = 0$. 当 Z 足够大,或胞腔半径足够小时,意味着 a_1 处或 a_2 处 $f_1 \approx 0$. 类似可推得当 Z 足够大时,任一顶点处都成立:$f_1 \approx 0, \cdots, f_{n-1} \approx 0$. 由于满足瓦尔拉斯条件

$$p_1 f_1 + \cdots + p_{n-1} f_{n-1} + p_n f_n = 0,$$

以及 $p_i > 0, i = 1, \cdots, n$,因此,$f_n$ 也将约等于0,从而全标号单纯形胞腔任一顶点坐标即为均衡价格的近似值,均衡产量也就相应可求得.

4.2 二要素多部门模型均衡点的求解

近几十年来一般均衡理论在实际中得到广泛的应用,并形成应用一般均衡(applied general equilibrium 或简称 AGE)理论体系. 应用一般均衡分析注重构造具体模型的技术及求解均衡点的算法. 由于应用一般均衡分析要计算出均衡价格与

均衡产量的具体数值，因此也称之为**可计算一般均衡**(computational general equilibrium 或简称 CGE) **分析**. 在实际应用中可以依实际情况构造出各种复杂程度不同的 AGE 或 CGE 模型. 本节及以下几个小节介绍几种简单的 CGE 模型及其构模技术.

二要素多部门模型所能描述的经济系统如下：设经济系统有 n 个部门，每个部门生产 1 种产品，各种产品生产过程中只投入资本 K 与劳动工时 L 两种要素. 在给定时间周期内，资本与劳动工时总可供量为常数. n 种产品价格分别为 $p_1,\cdots,p_n$，劳动工时价格为工资率 w，资本价格为租金 r.

二要素多部门模型具有采集数据方便等优点. 下面介绍模型基本方程.

第 i 种产品生产函数为

$$Q_i = A_i[\delta_i L_i^{\sigma_i} + (1-\delta_i)K_i^{\sigma_i}]^{1/\sigma_i}. \tag{4-8}$$

其中，Q_i 为第 i 种产品的产出量；K_i 与 L_i 分别为投入的资本与劳动工时；A_i, δ_i, σ_i 为常数，$-\infty < \sigma_i < 1$.

生产者在(4-8)式生产函数约束下谋求利润最大时，可求出单位产出要素需求函数为

$$\begin{aligned} l_i &= \frac{L_i}{Q_i} = l_i(r,w) \\ &= \frac{\left(\dfrac{w}{\delta_i}\right)^{1/(\sigma_i-1)}}{A_i\left[\delta_i\left(\dfrac{w}{\delta_i}\right)^{\sigma_i/(\sigma_i-1)} + (1-\delta_i)\left(\dfrac{r}{1-\delta_i}\right)^{\sigma_i/(\sigma_i-1)}\right]^{1/\sigma_i}}, \end{aligned} \tag{4-9}$$

$$\begin{aligned} k_i &= \frac{K_i}{Q_i} = k_i(r,w) \\ &= \frac{\left(\dfrac{r}{1-\delta_i}\right)^{1/(\sigma_i-1)}}{A_i\left[\delta_i\left(\dfrac{w}{\delta_i}\right)^{\sigma_i/(\sigma_i-1)} + (1-\delta_i)\left(\dfrac{r}{1-\delta_i}\right)^{\sigma_i/(\sigma_i-1)}\right]^{1/\sigma_i}}, \end{aligned} \tag{4-10}$$

其中，l_i 与 k_i 分别为单位产出劳动工时需求函数与资本需求函数.

产品价格 p_i 依投入的成本定价，即

$$p_i = wl_i + rk_i. \tag{4-11}$$

设有 m 个消费者，第 j 个消费者效用函数为

$$U_j = [a_{1j}^{1/\phi_j}x_{1j}^{(\phi_j-1)/\phi_j} + \cdots + a_{nj}^{1/\phi_j}x_{nj}^{(\phi_j-1)/\phi_j}]^{\phi_j/(\phi_j-1)}, \tag{4-12}$$

其中，x_{ij} 为第 j 个消费者享受到的第 i 种消费品数量，a_{ij} 及 ϕ_j 均为常数.

若第 j 个消费者的收入为 M_j 并全用于消费支出，则依效用最大法则可求出需求函数为

$$x_{ij} = \frac{a_{ij}p_i^{-\phi_j}M_j}{a_{1j}p_1^{1-\phi_j} + \cdots + a_{nj}p_n^{1-\phi_j}}. \tag{4-13}$$

设第 j 个消费者拥有资本为 K_j^s，同时他付出劳动工时为 L_j^s，那么他的收入

$$M_j = wL_j^s + rK_j^s. \tag{4-14}$$

下面介绍求解二要素多部门模型均衡点的计算方法与步骤.

第 1 步：计算精度的确定.

本模型是求均衡时 r 与 w 的值，因此标准单纯形是 r-w 平面上一条线段. 将该线段再划分成小线段(即剖分)，若取较大的 Z 值，则剖分越细，计算精度越高.

第 2 步：起始胞腔的选择.

起始胞腔两个顶点坐标写成矩阵形式为

$$\begin{bmatrix} 0 & 1/Z \\ 1 & (Z-1)/Z \end{bmatrix}, \tag{4-15}$$

其中，第 1 列所对应的顶点在标准单纯形边界上，第 2 列所对应的顶点在其内部.

第 3 步：胞腔顶点标号的计算.

(1) 将胞腔顶点坐标值作为 r 与 w 的初值，即 $r = 1/Z, w = (Z-1)/Z$.

(2) 当确定 r 与 w 值之后，按(4-9) 式与(4-10) 式计算 $l_i(r,w)$ 与 $k_i(r,w)$.

(3) 当计算出 l_i 与 k_i 之后，按(4-11) 式计算产品价格 p_i.

(4) 依(4-13) 式与(4-14) 式计算 x_{ij}，第 i 种产品总需求量 x_i 是全社会 m 个消费者对第 i 种消费品需求量之和. 即

$$x_i = \sum_{j=1}^{m} x_{ij}, \quad i = 1,\cdots,n. \tag{4-16}$$

(5) 第 i 种产品产出量 Q_i 在均衡时应等于需求量 x_i，即 $Q_i = x_i, i = 1,\cdots,n$.

(6) 计算要素需求量 L 与 K. 第 i 个生产者要素需求为

$$\begin{cases} L_i = l_i x_i = l_i Q_i, \\ K_i = k_i x_i = k_i Q_i. \end{cases}$$

全体生产者的要素需求量为

$$\begin{cases} L = L_1 + \cdots + L_n, \\ K = K_1 + \cdots + K_n. \end{cases}$$

(7) 计算要素总供给量为

$$L^s = L_1^s + \cdots + L_m^s, \quad K^s = K_1^s + \cdots + K_m^s.$$

(8) 计算要素市场供求差额为

$$E_l = L - L^s,$$
$$E_k = K - K^s.$$

(9) 计算顶点标号.

当顶点在标准单纯形内部时，标号规则如下：

若 $E_k > 0$，则顶点标号为 1；

若 $E_k \leqslant 0, E_l \geqslant 0$，则顶点标号为 2；

当顶点在标准单纯形边界时，如(4-15) 式所示矩阵的第 1 列，标号为该列先出现零的行号.

第 4 步：胞腔的转移.

胞腔的转移规则按(4-7) 式进行. 比如在(4-15) 式的起始胞腔中，第 1 列对应的顶点的标号为 1，若第 2 列对应的顶点的标号为 2，则找到全标号单纯形胞腔. 若

第2列对应的顶点的标号为1，则去掉第1列，即令 $j=1$，换上的第1列数值由(4-7)式计算为

$$b_1^1 = b_0^2 + b_0^2 - b_0^1 = 2\times b_0^2 - b_0^1$$
$$= 2\times\begin{bmatrix}1/Z\\(Z-1)/Z\end{bmatrix} - \begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}2/Z\\(Z-2)/Z\end{bmatrix},$$

这样起始胞腔便转移到新胞腔，新胞腔所对应的顶点坐标由如下矩阵描述：

$$\begin{bmatrix}2/Z & 1/Z\\(Z-2)/Z & (Z-1)/Z\end{bmatrix}.$$

换上新顶点之后，再按上述规则计算它的标号，直至找到全标号单纯形胞腔为止.

第5步：均衡点的计算.

当找到全标号单纯形胞腔后，它有2个顶点，分别设为 A 点与 B 点，标号分别为 $l(A)=1, l(B)=2$. 按上述标号规则，$l(A)=1$ 意味着在 A 处 $E_k>0$；$l(B)=2$ 意味着在 B 处 $E_k\leqslant 0$. 因此，在 A 与 B 附近必存在1点使得 $E_k=0$；当 Z 足够大，胞腔半径足够小时，在 A 处可认为 $E_k\approx 0$. 可以证明两要素多部门模型满足瓦尔拉斯条件

$$rE_k + wE_l = 0,$$

其中，$E_k\approx 0$ 意味着 $E_l\approx 0$. 因此全标号单纯形胞腔任一顶点坐标值都可以当做市场均衡时的 r 与 w 值. 当找到均衡点的 r 与 w 之后，产品价格 p_i 及供求平衡量也就都相应计算出来了.

两要素多部门模型存在一些缺点，如：① 租金收入 rK' 全部用于消费，没有考虑固定资产投资问题；② 没有考虑税收收入与公共消费支出问题；③ 没有考虑关税与进出口问题，等等. 以上缺点可以通过构造更复杂的模型来不同程度地克服.

4.3 考虑税收政策的二要素多部门模型均衡点的求解

本节将上一节二要素多部门模型推广到考虑税收政策的情况中去.

首先，考虑租金所得税. 设租金所得税税率为 τ_k，税基为资本出租者出租资本实际所得，即租金 r，因此生产者每使用1单位资本实际所付租金为 $r\times(1+\tau_k)$.

其次，考虑工资所得税. 设工资所得税税率为 τ_l，税基为劳动者工资收入 w，因此生产者每使用1单位劳动工时应付工资为 $w\times(1+\tau_l)$.

设经济系统有 n 种产品，每种产品只使用两种要素，其生产函数同(4-8)式所示. 由于考虑资本所得税与工资所得税后，生产者面对的资本价格与劳动工时价格分别为 $r(1+\tau_k)$ 与 $w(1+\tau_l)$，因此，单位产出要素需求函数只须将(4-9)式与(4-10)式中的 r 与 w 相应改为 $r(1+\tau_k)$ 与 $w(1+\tau_l)$ 即可. 即

$$l_i = L_i/Q_i = l_i(r(1+\tau_k), w(1+\tau_l)),$$
$$= \frac{\left(\dfrac{w(1+\tau_l)}{\delta_i}\right)^{1/(\sigma_i-1)}}{A_i\left[\delta_i\left(\dfrac{w(1+\tau_l)}{\delta_i}\right)^{\sigma_i/(\sigma_i-1)} + (1-\delta_i)\left(\dfrac{r(1+\tau_k)}{1-\delta_i}\right)^{\sigma_i/(\sigma_i-1)}\right]^{1/\sigma_i}},$$

$$k_i = K_i/Q_i = k_i(r(1+\tau_k), w(1+\tau_l))$$

$$= \frac{\left(\frac{r(1+\tau_k)}{1-\delta_i}\right)^{1/(\sigma_i-1)}}{A_i\left[\delta_i\left(\frac{w(1+\tau_l)}{\delta_i}\right)^{\sigma_i/(\sigma_i-1)}+(1-\delta_i)\left(\frac{r(1+\tau_k)}{1-\delta_i}\right)^{\sigma_i/(\sigma_i-1)}\right]^{1/\sigma_i}}. \tag{4-17}$$

第 i 种产品应税价格(即税基)p_i,应等于生产者每销售1单位该种产品的实际所得.按成本定价它由下式计算:

$$p_i = w\times(1+\tau_l)\times l_i(r(1+\tau_k), w(1+\tau_l)) + r\times(1+\tau_k)\times k_i(r(1+\tau_k), w(1+\tau_l)). \tag{4-18}$$

现考虑第 i 种产品销售税:销售税税率为 τ_i,税基为 p_i.销售1单位第 i 种产品应纳销售税 $\tau_i\times p_i$.这样,消费者面对的市场价为 $\tilde{p}_i = p_i\times(1+\tau_i)$.

在本模型中共有三种税:资本所得税、工资所得税、销售税.三种税的税收总额

$$T = r\tau_k K^s + w\tau_l L^s + \sum_{i=1}^{n} p_i\tau_i Q_i, \tag{4-19}$$

其中,K^s 与 L^s 为资本与劳动总可供量,Q_i 为第 i 种产品总产出量.

设税收支出总额为 $\tilde{T}$,在财政收支平衡时它应等于税收收入总额 T.由于本模型没有考虑公共消费,因此税收收入全部转移支付给消费者作为个人消费支出.

设第 j 种人获得的转移支付额为 $\tilde{T}\times\delta^j$,共有 m 种人,$\delta^1+\cdots+\delta^m=1$.第 j 种人收入 M_j 为劳动收入 wL_j^s,加上资本收入 rK_j^s,再加上税收转移支付收入.即

$$M_j = wL_j^s + rK_j^s + \tilde{T}\ \delta^j. \tag{4-20}$$

如果第 j 种人对各种产品效用函数可用(4-12)式来表达,由于消费者面对的是含税市场价格 $\tilde{p}_i$,因此他对第 i 种产品需求量 x_{ij} 为

$$x_{ij} = \frac{a_{ij}\tilde{p}_i^{-\phi_j}M_j}{a_{1j}\tilde{p}_1^{1-\phi_j}+\cdots+a_{nj}\tilde{p}_n^{1-\phi_j}}. \tag{4-21}$$

第 i 种产品产出量 Q_i 应等于全体消费者消费需求,即

$$Q_i = x_{i1}+\cdots+x_{im}. \tag{4-22}$$

劳动工时总供给量为每种人付出的劳动工时之和,即

$$L^s = L_1^s+\cdots+L_m^s. \tag{4-23}$$

劳动工时总需求量 L^D 为各生产者对劳动的要素投入需求之和,即

$$L^D = L_1+\cdots+L_n. \tag{4-24}$$

劳动工时供求差额 E_l 为

$$E_l = L^D - L^s. \tag{4-25}$$

资本总供给量 K^s 为每种人拥有资本的总和,即

$$K^s = K_1^s + \cdots + K_m^s.$$

资本总需求量 K^D 为各生产者对资本的要素投入需求之和，即

$$K^D = K_1 + \cdots + K_n.$$

资本供求差额 E_k 为

$$E_k = K^D - K^s. \qquad (4\text{-}26)$$

考虑税收时二要素多部门模型市场均衡点计算步骤如下：

第1步：计算精度的确定与起始胞腔的选择.

在 r-w-$\tilde{T}$ 三维空间 $\mathbf{R}^3$ 中对标准单纯形进行剖分，起始胞腔三个顶点坐标由(4-4)式所示矩阵表示($n = 3$). Z 值越大剖分越细，则计算精度相对越高.

第2步：胞腔顶点标号的计算.

若胞腔顶点在标准单纯形边界上，则按(4-5)式规则标号. 若胞腔顶点在标准单纯形内部，则按如下步骤计算它的标号：

(1) 将胞腔顶点坐标值作为 r, w 与 $\tilde{T}$ 的初值.

(2) 按(4-17)式计算 l_i 与 k_i.

(3) 按(4-18)式计算 p_i，再计算 $\tilde{p}_i = p_i \times (1 + \tau_i)$.

(4) 按(4-20)式计算消费支出.

(5) 按(4-21)式计算第 j 个消费者对第 i 种产品需求量. 再按(4-22)式计算全体消费者对该种产品总需求量 Q_i，它即为该种产品总产出量.

(6) 按(4-19)式计算税收总额 T.

(7) 按(4-25)式计算劳动工时供求差额 E_l，按(4-26)式计算资本供求差额 E_k，再计算税收入 T 与支出 $\tilde{T}$ 之间差额 $E_g = T - \tilde{T}$.

(8) 如果顶点坐标分量分别对应 $r, w, \tilde{T}$ 的数值，则按如下标号规则标号：若 $E_k > 0$，则标号为1；若 $E_k \leqslant 0, E_l > 0$，则标号为2；若 $E_k \leqslant 0, E_l \leqslant 0, E_g \geqslant 0$ 则标号为3.

第3步：胞腔的转移.

如果没有找到全标号单纯形胞腔，则按(4-7)式规则从原有胞腔转移到新胞腔，直至找到全标号单纯形胞腔为止.

第4步：均衡点的计算.

当在标准单纯形内部找到一个全标号单纯形胞腔后，该胞腔任一顶点坐标都可近似作为 $r, w, \tilde{T}$ 的均衡点近似值. 这是因为可以证明本模型满足如下瓦尔拉斯条件：

$$rE_k + wE_l + E_g = 0.$$

全标号单纯形胞腔意味着在胞腔任一顶点处成立 $E_k \approx 0, E_l \approx 0$，再依上式有 $E_g \approx 0$.

一旦求出 $r, w, \tilde{T}$ 的均衡值,市场价格、产量、需求量、税收收入总量等数值便可相应求出.

对二要素多部门模型,不难将它再进一步推广到多部门且考虑税收的情况.

4.4　考虑国际贸易与关税政策的可计算一般均衡模型

二要素多部门模型可推广为考虑国际贸易及资本所得税、工资所得税与产品销售税,并考虑关税的模型.这时变量将更多,模型也将更加复杂.由于这里的主要目的在于讨论可计算一般均衡(CGE)模型构模技术,本节模型将不考虑资本所得税、工资所得税及产品销售税,仅考虑关税.读者不难将上一节方法与本节模型相结合,从而构造更复杂更完善的模型.下面给出模型基本方程.

设有 n 个部门,每个部门生产1种产品,称之为国产品.每种产品除国产品外,还相应有进口品.每种产品的生产可以投入资本、资源、劳动工时(可以包括简单劳动与复杂劳动)等各种要素.为叙述简洁起见,只考虑两个部门及资本与劳动两种要素.

生产函数为

$$x_i = z_i L_i^{\beta_i} K_i^{r_i} x_{1i}^{\alpha_{1i}} x_{2i}^{\alpha_{2i}} m_{1i}^{\phi_{1i}} m_{2i}^{\phi_{2i}}, \tag{4-27}$$

其中,$i = 1$ 表示工业部门,$i = 2$ 表示农业部门;

x_i 为第 i 部门总产出量;

L_i 为第 i 部门投入的劳动工时;

K_i 为第 i 部门投入的固定资本;

x_{1i} 为第 i 部门消耗的国产工业品数量;

x_{2i} 为第 i 部门消耗的国产农业品数量;

m_{1i} 为第 i 部门消耗的进口工业品数量;

m_{2i} 为第 i 部门消耗的进口农业品数量;

$z_i, \beta_i, r_i, \alpha_{1i}, \alpha_{2i}, \phi_{1i}, \phi_{2i}$ 为给定参数,并且满足

$$\beta_i + r_i + \alpha_{1i} + \alpha_{2i} + \phi_{1i} + \phi_{2i} = 1.$$

在生产者谋求利润最大时,可求出单位产出时的各投入要素需求量分别为

$$\begin{cases} l_i = L_i/x_i = \beta_i p_i/w, \\ k_i = K_i/x_i = r_i p_i/r, \\ x_{1i}/x_i = \alpha_{1i} p_i/p_1, \\ x_{2i}/x_i = \alpha_{2i} p_i/p_2, \\ m_{1i}/x_i = \phi_{1i} p_i/q_1, \\ m_{2i}/x_i = \phi_{2i} p_i/q_2. \end{cases} \tag{4-28}$$

其中,w, r, p_1, p_2, q_1, q_2 分别为劳动工时工资率、资本租金、国产工业品价格、国产农业品价格、进口工业品价格、进口农业品价格.

国产品价格 p_i 与投入要素价格有关，由下式计算：

$$p_i = \frac{1}{z_i}\left(\frac{w}{\beta_i}\right)^{\beta_i}\left(\frac{r}{r_i}\right)^{r_i}\left(\frac{p_1}{\alpha_{1i}}\right)^{\alpha_{1i}}\left(\frac{p_2}{\alpha_{2i}}\right)^{\alpha_{2i}}\left(\frac{q_1}{\phi_{1i}}\right)^{\phi_{1i}}\left(\frac{q_2}{\phi_{2i}}\right)^{\phi_{2i}}. \tag{4-29}$$

下面分析消费者对国产品与进口品的最终消费需求函数.

设 D_1 为国产工业品消费需求量，D_2 为国产农业品消费需求量，M_1 为进口工业品消费需求量，M_2 为进口农业品消费需求量. 当消费者购入 D_1, D_2, M_1, M_2 数量的各种产品之后，便得到相应效用 U，效用函数 U 可表示为

$$U = [\delta_1 D_1^{\rho_1} + (1-\delta_1)M_1^{\rho_1}]^{\varepsilon_1/\rho_1} \times [\delta_2 D_2^{\rho_2} + (1-\delta_2)M_2^{\rho_2}]^{\varepsilon_2/\rho_2}, \tag{4-30}$$

其中，$\varepsilon_1 + \varepsilon_2 = 1$.

当消费者消费支出总额为 Y 时，预算约束方程由下式描述：

$$p_1 D_1 + p_2 D_2 + q_1 M_1 + q_2 M_2 = Y. \tag{4-31}$$

在上述预算约束之下求效用最大，可得到如下需求函数表达式：

$$\begin{cases} D_1 = \varepsilon_1 p_1^{-\sigma_1}\delta_1^{\sigma_1}Q_1^{\sigma_1-1}Y; \\ D_2 = \varepsilon_2 p_2^{-\sigma_2}\delta_2^{\sigma_2}Q_2^{\sigma_2-1}Y; \\ M_1 = \varepsilon_1 q_1^{-\sigma_1}(1-\delta_1)^{\sigma_1}Q_1^{\sigma_1-1}Y; \\ M_2 = \varepsilon_1 q_2^{-\sigma_2}(1-\delta_2)^{\sigma_2}Q_2^{\sigma_2-1}Y. \end{cases} \tag{4-32}$$

其中

$$Q_i = [\delta_i^{\sigma_i} p_i^{1-\sigma_i} + (1-\delta_i)^{\sigma_i} q_i^{1-\sigma_i}]^{1/(1-\sigma_i)}.$$

国产品出口需求量 E_i 与第 i 种产品价格 p_i 及汇率 e 有关，e 为人民币汇率，即 1 美元兑换 e 元人民币. 出口需求函数可以简单地由下式表达：

$$E_i = g_i\left(\frac{p_i}{e}\right)^{\eta_i}, \quad i = 1,2, \quad g_i > 0, \quad \eta_i < 0. \tag{4-33}$$

国产品产出量 x_i 用于中间投入 $x_{i1} + x_{i2}$ 以及消费需求 D_i 加上出口需求 E_i，因此有如下平衡方程式：

$$\boldsymbol{x} = \boldsymbol{A}(p)\cdot\boldsymbol{x} + \boldsymbol{D} + \boldsymbol{E}, \tag{4-34}$$

其中

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix},$$

中间消耗系数阵

$$\boldsymbol{A}(p) = \begin{bmatrix} x_{11}/x_1 & x_{12}/x_2 \\ x_{21}/x_1 & x_{22}/x_2 \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12}p_2/p_1 \\ \alpha_{21}p_1/p_2 & \alpha_{22} \end{bmatrix}.$$

(4-34) 式是通常线性列昂锡夫静态投入产出模型的推广.

下面讨论国际贸易与国际收支平衡问题.

设第 i 种产品国际市场价为 p_{w_i}（按美元计），按人民币计的国际市场价为 ep_{w_i}. 设该种产品关税税率为 τ_i，税基为该种产品国内市场价 q_i，因此进口 1 单位该种产品应征收关税 $\tau_i q_i$ 元人民币. 国际市场价 ep_{w_i} 加上关税 $\tau_i q_i$ 应等于该种产品国内市场价 q_i，即

$$ep_{w_i} + \tau_i q_i = q_i \quad 或 \quad q_i = ep_{w_i}/(1-\tau_i). \tag{4-35}$$

国际收支平衡时应成立下式：

$$进口需汇 = 出口创汇 + 关税收入. \tag{4-36}$$

其中

$$出口创汇 = p_1E_1 + p_2E_2,$$

$$进口需汇 = q_1M_1 + q_2M_2 + q_1(m_{11} + m_{12}) + q_2(m_{21} + m_{22}),$$

$$关税收入 = \tau_1q_1M_1 + \tau_2q_2M_2 + \tau_1q_1(m_{11} + m_{12}) + \tau_2q_2(m_{21} + m_{22}).$$

当国际收支不平衡时，存在差额 E_f，

$$E_f = 进口需汇 - 出口创汇 - 关税收入. \tag{4-37}$$

若 $E_f > 0$，表示国际收支存在逆差；

若 $E_f < 0$，表示国际收支为顺差.

市场均衡点的求解步骤如下：

(1) 给定 $p_{w_i}, \tau_i, K^s, L^s$ 值，它们为已知常量.

(2) 选 w, r, e 的初值.

求解均衡点的任务就是要求出 w, r, e 的均衡解，使系统的资本市场、劳动市场与国际收支达到平衡.

在 w-r-e 三维空间中，标准单纯形的起始胞腔为

$$\begin{bmatrix} 0 & 1/Z & 1/Z \\ 1/Z & 0 & 1/Z \\ (Z-1)/Z & (Z-1)/Z & (Z-2)/Z \end{bmatrix},$$

其中，第 3 列顶点坐标作为 w, r, e 的初值.

(3) 按(4-35) 式计算 q_i.

(4) 按(4-29) 式计算 p_i.

(5) 按(4-33) 式计算 E_i.

(6) 按(4-28) 式计算 k_i, l_i，确定 $A(p)$.

(7) 按(4-28) 式确定进口品中间消耗系数阵 $B(p,q)$，

$$B(p,q) = \begin{bmatrix} m_{11}/x_1 & m_{12}/x_2 \\ m_{21}/x_1 & m_{22}/x_2 \end{bmatrix} = \begin{bmatrix} \phi_{11}p_1/q_1 & \phi_{12}p_2/q_1 \\ \phi_{21}p_1/q_2 & \phi_{22}p_2/q_2 \end{bmatrix}.$$

(8) 计算国产品产出量 x_i.

首先，可以证明消费总支出额

$$\begin{aligned} Y &= wL^s + rK^s + 关税收入 \\ &= wL^s + rK^s + \tau_1q_1M_1 + \tau_2q_2M_2 + \\ &\quad [\tau_1q_1, \tau_2q_2]B(p,q)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \end{aligned} \tag{4-38}$$

然后，按(4-34)，(4-38) 式及(4-32) 式中 M_i 表达式，可联立求出 x_i 值.

(9) 计算资本市场供求差额 E_k.

$$E_k = K_1^D + K_2^D - K^s = k_1x_1 + k_2x_2 - K^s. \tag{4-39}$$

(10) 计算劳动市场供求差额 E_l.

$$E_l = L_1^D + L_2^D - L^s = l_1x_1 + l_2x_2 - L^s. \tag{4-40}$$

(11) 按(4-37)式计算国际收支差额 E_f.

(12) 顶点标号的计算与胞腔的转移.

如果顶点坐标有0分量,首先出现0的行号为顶点标号.若顶点在标准单纯形内部,则按下面规则标号:

若 $E_l > 0$,则该顶点标号为1;

若 $E_l \leqslant 0, E_k > 0$,则该顶点标号为2;

若 $E_l \leqslant 0, E_k \leqslant 0, E_f \geqslant 0$ 则该顶点标号为3.

如果没有找到全标号单纯形胞腔,则按(4-7)式的方法进行胞腔转移,直至找到全标号单纯形胞腔为止.

(13) 瓦尔拉斯条件与均衡点的求解.

可以证明本系统满足如下瓦尔拉斯条件:

$$wE_l + rE_k + E_f = 0. \tag{4-41}$$

当 Z 足够大,即胞腔半径足够小时,(4-41)式可以保证全标号单纯形胞腔任一顶点坐标都可以作为 w, r, e 的均衡解.一旦求出了 w, r, e 的均衡解,那么就确定了市场供求平衡时国产品产量及其价格、进口品数量及其价格、消费者需求量等各项数值.

4.5 可计算一般均衡模型的进展

以上几节给出了几个典型的可计算一般均衡模型,分析了构模基本技术与相应算法.应当指出,这些模型尚有许多不足.例如,

(1) 在实际经济中,税收收入一方面用于公共消费(国防、环境保护、医疗卫生等),另一方面用于公共投资(水利、交通等),但以上模型没有考虑公共消费与公共投资,也没有考虑教育、科研与人口增长等.

(2) 人们的收入并不一定都用于消费,而是部分用于消费,部分用于储蓄.如果考虑消费者的储蓄行为,则模型中还应加进利率因素,即以上模型将被扩充至包含货币供应与货币需求的货币市场中去.

(3) 以上模型是完全竞争模型,没有考虑垄断行为及信息交流的不充分性.

(4) 以上模型为确定性模型,没有考虑各种随机因素.

(5) 没有考虑按劳分配问题,等等.

近几十年来,可计算一般均衡理论与应用得到迅速的发展,各国的经济学家构造出了不同类型的CGE模型,以上所述的不足之处都在不同程度上得到克服与解决.

今后CGE理论与应用进展有如下一些特点:

(1) 考虑因素将更加全面,系统中将包括产品市场、劳动市场、资源市场、资本市场、货币市场,还要考虑教育、科研、环境保护、国防安全等.

(2) 包括的变量更多,不仅考虑一国的经济增长,还要考虑多国间的经济贸易

与经济交流问题.

(3) 不仅考虑如何求静态平衡点,还要考虑如何求解协调增长与最优增长问题,以及相应的货币政策、财政政策的设定.

(4) 在计算方法与应用软件上将不断改善,计算速度更快,方法更简洁.

目前 CGE 理论或数理经济学理论与实践正处在快速发展阶段.

5　数理经济学的其他基本研究方向

5.1　概　　述

数理经济学是一门内容极为广泛的学科,但归纳起来有如下 5 个基本问题.

第 1 个问题:市场均衡点的存在性与唯一性.

"供给、需求、平衡"是经济学最基本、最核心的 6 个字.数理经济学的首要任务是给出供给函数、需求函数的定量描述,并找出供求平衡的较准确位置.本篇正是着重讨论了产品市场的供求平衡问题.应当指出,数理经济学几位著名代表人物瓦尔拉斯、阿罗、德布鲁、列昂锡夫等人虽然在解决第 1 个问题方面做出了卓越贡献,并多次获得诺贝尔奖,但是离第 1 个问题的完美解决尚有很大距离.因为他们主要是在无国际贸易情况下讨论了产品市场的供求平衡,而现实的任务是要在考虑国际贸易及不完全竞争市场情况下,求解包括产品市场、资本市场、劳动市场、资源市场、货币市场、信息市场、技术市场等众多市场的供求平衡问题,在理论上甚至要考虑包含无穷多种产品的市场供求平衡问题.随着理论与实践的深入,在求解市场均衡点方面又有许多新成果问世,继瓦尔拉斯、阿罗、德布鲁、列昂锡夫等人之后,又出现了许多新的代表人物.

第 2 个问题:市场运动均衡点的稳定性.

由于各种自然的或人为的原因,经济系统将偏离供求平衡点并产生经济波动.这种经济波动是越来越剧烈,还是逐渐趋于平稳?这个问题属于经济系统运动稳定性分析范畴.亚当·斯密在 200 多年前就指出:每日每时有许多生产者生产各种产品,又有许多消费者购买这些产品.虽然不存在一些具体的人在操纵市场的供给与需求,而市场却最终能自动达到供求平衡位置.亚当·斯密的这一论述便是朴素的经济系统运动稳定性概念.亚当·斯密还指出,之所以现实经济能自动到达供求平衡点,是因为有一个看不见的人伸出一只看不见的手在操纵市场.这"看不见的手"是什么呢?现代经济理论指出,它就是市场的价格机制.在现代控制理论中有一个新分支叫**鲁棒(robust)调节理论**.经济学家已指出鲁棒调节理论与亚当·斯密"看不见的手"论述之间有紧密联系.现代控制理论中的鲁棒调节理论与**稳定性分析理论**,是分析经济系统运动稳定性的有力工具.由于经济系统的大规模性与复杂性,全面讨论经济系统运动的稳定性是一件复杂而艰深的工作,这个领域有许多工作等待人们去探索与发现.

第3个问题:经济系统运动的合理性与经济系统目标的设定.

经济学的定义是:利用有限资源合理安排生产,生产出来的产品在消费者中进行合理分配,以达到人类现在和将来的最大满足.数理经济学首要任务是:给出人类满足的定量描述,达到人类的最大满足.简单地说,人类的最大满足就是要达到公平与有效益的境界."效益"指的是物质的极大丰富以及环境、健康等其他许多因素.然而,什么叫"公平"?这个问题资本主义经济学家与社会主义经济学家有不同的判断准则.资本主义经济学家认为**按资分配**是公平的境界,而社会主义经济学家则认为**按劳分配**是公平的境界.经济系统的目标设定属规范经济学范畴,所谓规范经济学就是要给出什么叫"好"的价值判断准则.关于第3个问题,当前经济学有极为广泛与丰富的成果,由于篇幅所限本篇基本上没有涉及并展开论述.下面列举几位代表人物及其相应成果.

代表人物之一:马克思.**马克思的劳动价值论**无疑是经济学历史上最伟大、最重要的贡献之一.马克思提出了产品的价值、剥削、剩余价值、按劳分配等非常重要的基本概念,为制定社会主义及共产主义的目标奠定了基础.马克思认为,社会主义目标是物质较大丰富并达到按劳分配境界;共产主义目标是物质极大丰富并达到各尽所能按需分配的境界.这也就给出了社会主义与共产主义效益与公平的基本定义.

代表人物之二:森岛(Morishima).日本的森岛等人首先力图将马克思劳动价值论定量化.在1950～1960年,森岛等人利用线性多部门生产函数理论对马克思劳动价值论作了定量描述,给出了计算凝结在一种产品中的社会必要劳动时间的方法,给出了剥削及剩余价值的定量描述.尽管森岛等人的工作尚存在争议,但他们在沟通马克思经济学与西方的微观、宏观经济学方面起到极其重要的作用.

代表人物之三:帕雷托(V. Pareto)是意大利经济学家."**帕雷托最优**"是经济学中最基本与最重要的概念之一.无论社会主义的市场经济还是资本主义市场经济,都要达到消费品与生产资料的帕雷托最优配置.但是帕雷托最优解一般有无穷多个,可把既符合帕雷托最优,又符合马克思按劳分配准则的配置叫做"**马克思最优境界**".我国经济学工作者给出了"马克思最优境界"定量化的初步描述.马克思最优境界,是社会主义市场经济的目标境界.

除以上几个代表人物外,阿罗等人所创立的福利经济学也与这个领域密切相关.

第4个问题:市场运动的能控性与宏观经济调控政策的设计.

当给出经济系统目标的定量化描述之后,还要明确什么是经济系统的政策变量或调控变量,以及应当采用什么样的调控政策让系统沿最优轨道前进并到达理想的目标境界.一般地说,经济系统有两大政策:货币政策与财政政策.财政政策的内容是:增值税、企业所得税、个人所得税等各种税的最优税率是多少?税收收入占GNP合理比重是多少?应当怎样合理支配税收收入(即公共消费与公共投资合适比例是多少?具体地说,国防、环境保护、科研、教育、医疗卫生、交通、水利等合理支出是多少)?此外,老龄人口的社会保险的财政补贴,富裕地区与人群的税收收入对相对贫困地区与人群的财政转移支付等问题都属于财政政策所涉及的范畴.货币政

策的内容是:货币供应量应多大?货币供应量如何与经济增长相适应,从而有效抑制过高的通货膨胀率?此外,在银行与企业的关系中,银行要决定可以向什么样的企业或什么类型的行业贷款、贷款额度多大,也应当归入货币政策的范畴.当采用一定的宏观经济政策之后,这些政策能否起作用,即能否有效地将现实系统调控到目标轨道上,这便是经济系统的能控性问题.现代控制理论中最优控制理论分支以及庞得里亚金(Poutryagin)极大值原理等知识是分析经济系统能控性的有力工具.由于经济系统一般为大规模非线性动态随机灰色参数系统,因此讨论经济系统能控性问题是一件难度极大的工作.目前在经济学中关于经济系统能控性与最优轨道设计方面最重要的成果,是由萨缪尔森(Samuelson)等人所创建的快车道(Turnpike)定理.20世纪50年代萨缪尔森等人应用线性规划设计线性多部门模型的最优轨道时,得到了线性多部门模型的快车道定理.快车道定理认为,长期来讲,线性经济系统应当运行在平衡增长轨道上(平衡增长轨道又称为冯·诺伊曼射线),如果经济系统不在平衡增长轨道上,就应尽快调控到该轨道上,因此平衡增长轨道又称为快车道.应当指出,对大规模线性经济系统可以有效地求解最优增长轨道或快车道,但对大规模非线性动态经济系统来讲,最优轨道的计算与求解尚在不断地探索之中.

第5个问题:经济系统在一定时间内到达合理位置的能达性.

如果当前经济系统不在合理位置,那么要花多少时间才能调控到合理位置?此外,发展中国家多少年才能赶上发达国家?等等,都属于这个领域.近年来经济学家开始研究发展中国家对发达国家的追赶问题.这方面的研究也处在迅速发展中.

5.2 经济控制论的理论与应用

经济控制论强调的是,用系统论的思想与控制理论的方法来描述经济学、数理经济学强调的是,用数学的方法来描述经济学.由于控制理论可以看作是应用数学的一个组成部分,因此经济控制论也可以看作是数理经济学的一个组成部分.如上所述,数理经济学有5个基本问题.目前习惯上把讨论解的存在性与求解方法、经济系统运动有效性与合理性称为数理经济学,而经济控制论侧重讨论稳定性、经济系统目标设定、能控性、一定时间内到达合理位置的能达性.因此可以说经济控制论就是数理经济学,而且它涉及到了数理经济学中更艰深更困难的部分.

依据描述经济系统运动的方程类型,可将经济系统分为:确定性系统、随机系统、决策系统、对策系统、线性系统、非线性系统、灰色参数系统、集中参数系统、分布参数系统、精确系统、模糊系统,等等.因此经济控制论可以有许多分支:确定性动态系统经济控制论、广义动态系统经济控制论、随机动态经济系统的分析与控制、非线性多部门经济系统分析与控制、经济系统的模糊控制理论等,所有这些分支也都可以归于数理经济学的范畴.对于经济控制论的基本内容本书另辟专篇介绍.

5.3　非线性经济系统的理论与应用

当用各种数学方程描述经济运动规律时，除了得到线性的微分方程组外，往往还会得到连续时间或离散时间的非线性微分方程组或差分方程组. 因此非线性系统理论与方法对经济学来讲是基本的且是十分重要的.

一般地说，经济系统的数学模型可以归结为用微分方程组

$$\frac{dx}{dt} = f(x, u), \tag{5-1}$$

或差分方程组

$$x(t+1) = f(x(t), u(t)) \tag{5-2}$$

表示. 其中，x 为系统状态变量，u 为控制变量或政策变量.

对上述系统来讲，如下问题是基本的：

1. 平衡点的存在性

平衡点方程为

$$f(x, u) = 0,$$

本篇前面所着重讨论的一般均衡理论可归入这个范畴.

2. 平衡点的局部稳定性与全局稳定性

市场调节的稳定性分析可归入这个范畴.

3. 非线性经济系统的分支与混沌现象

有些经济系统由于政策变量的作用或现实的随机干扰，系统可能到达混沌状态. 非线性系统分支的研究可以通过多元函数 $f(x, u)$ 的泰勒展开式来进行分析.

4. 极限环与经济周期

所谓极限环是指系统围绕平衡点作周期运动. 在生态平衡等系统中往往存在极限环.

5. 大规模经济系统的最优控制

应用控制理论中庞得里亚金极大值原理来求解大规模经济系统的最优轨道，目前尚缺乏有效的算法. 但对一些特定的非线性多部门经济系统，可以根据其特殊规律找到有效的求解方法.

5.4　经济对策系统的理论与应用

当经济系统所有政策变量由一个决策者所掌握，并且该决策者将各个子目标综合成一个总目标时，称这种系统为决策系统. 如果经济系统有多个决策者，每个决策者各有自己的决策变量与目标变量，那么称这种系统为对策系统. 现实的经济系统是对策系统. 一般地说，现实经济有三种决策者：政府、消费者、生产者. 政府的决策变量为税率、税收总量、财政支出分配等，目标变量是全社会达到公平与有效的境界. 消费者决策变量为消费品购入量、劳动工时提供量等，目标变量为各自效用最大. 生产者决策变量为各产品产出量等，目标变量为各自利润最大等. 由此可

知，对策论在经济领域具有极其广泛与卓有成效的应用．有关这方面的文献也极为丰富．

总之，数理经济学包含着极为广泛和深刻的内容，本篇介绍的仅是其中最基本、最重要的一部分内容．

参考文献

1 潘吉勋，张顺明著．经济均衡的数学原理．长春：吉林大学出版社，1997．
2 李楚霖，林少宫著．微观经济的数理导引．武汉：华中工学院出版社，1985．
3 胡显佑，龚德恩著．线性经济模型及其数学方法．北京：中国人民大学出版社，1995．
4 王则柯著．单纯不动点算法基础．广州：中山大学出版社，1986．
5 潘介人著．数理经济．上海：上海交通大学出版社，1989．
6 张金水著．数理经济学——理论与应用．北京：清华大学出版社，1998．
7 (美)Woods J E 著．数理经济学．代定一，汪同三，秦朵译．北京：中国展望出版社，1987．
8 Arrow K J，Intriligator M D. Handbook of mathematical economics. vol.Ⅰ，Ⅱ，Ⅲ. Amsterdam: North-Holland，1985．
9 Takayama A. Mathematical economics. New Yourk: Cambridge University Press，1985．
10 Shoven J B，Whalley J. Applying general equilibrium. New York: Cambridge University Press，1992．

第3篇

金融数学

编　者　严加安
审校者　龚光鲁

目　录

引　言

金融数学是金融经济学的数学化，通过建立金融市场中证券价格和利率的期限结构的数学模型，研究风险资产（包括衍生金融产品和金融工具）的定价、对冲(hedging)策略和最优投资消费策略的构造以及风险管理．

金融数学的历史可以追溯到20世纪初．1900年法国概率学家巴施里埃(L. Bachelier)首次提出用布朗运动(Brownian motion)描述股票价格的变动．1952年马柯维茨(H. M. Markowtz)提出了用于投资分析和风险管理的均值-方差分析方法．1958年莫迪里亚尼(F. Modigliani)和米勒(W. Miller(等))从"套利推理"(arbitrage arguments)出发对公司财务理论进行了研究，得出了"在无税收和公司不破产前提下，公司的价值与公司的资本结构无关"这一结论(被称为MM定理)．60年代中期，在马柯维茨的均值-方差分析基础上，夏普(W. F. Sharpe)、林特纳(J. Lintner)和毛新(J. Mossin)进一步发现在竞争均衡的市场中，每种风险资产的预期收益率与市场投资组合(market portfolio)的风险报酬之间有一个线性关系，这就是著名的资本资产定价模型(CAPM)．马柯维茨的均值-方差分析和夏普等人提出的资本资产定价模型后来被誉为"华尔街的第一次革命"．马柯维茨和夏普获得1990年度诺贝尔经济学奖，同时获奖的还有前面提到的米勒．

衍生金融产品（如期权）的合理定价是金融数学研究的中心问题之一．1973年，布莱克(F. Black)和索尔斯(M. Scholes)基于套利推理，从构造一个对冲交易策略出发，导出了著名的(Black-Scholes)期权定价公式．同年，默顿(R. C. Merton)对布莱克-索尔斯(Black-Scholes)模型和定价公式作了完善和多方面的推广．由他们开创的期权定价理论(OPT)被誉为"华尔街的第二次革命"．索尔斯和默顿因此荣获1997年度诺贝尔经济学奖(布莱克于1995年英年早逝，未能分享此殊荣)．

本篇主要介绍金融数学的核心——资本资产定价模型和期权定价理论的基础性内容．因篇幅所限，不介绍属于公司财务理论的MM定理，对动态投资组合理论也只作粗略的介绍．

1　静态投资组合理论

1.1　投资组合的选择理论

1.1.1　收益率与风险

证券投资的收益由两个部分组成：一是本期收入（如债券利息或股票股息）；二

是资本利得或损失，即由于证券价格的变动产生的价差．收益额与投资额的比率称为**收益率**．由随机因素引起收益率不确定的证券称为风险证券．在美国，一般把联邦政府发行的短期国库券当做无风险证券，它的收益率是预先确定的．所谓风险，通常是指导致投资损失的可能性．如果把风险证券收益率看成一个随机变量，则它的数学期望称为**预期收益率**(expected rate of return)，它的方差或标准差可以作为证券风险的一种度量指标．

对单期间交易的证券市场，假定市场中有 N 种风险证券及1种无风险证券，在期间末第 i 种风险证券的收益率为 r_i，无风险证券的收益率为一正常数 r_f，则可分别用 $\boldsymbol{r}=(r_1,\cdots,r_N)^{\mathrm{T}}$ 和 $\boldsymbol{e}=(e_1,\cdots,e_N)^{\mathrm{T}}$ 表示风险证券的收益率向量和预期收益率向量(其中 $e_i=E[r_i]$，E 为期望算子)，用 $\boldsymbol{V}$ 表示 $\boldsymbol{r}$ 的协方差阵，即 $\boldsymbol{V}=E[(\boldsymbol{r}-\boldsymbol{e})(\boldsymbol{r}-\boldsymbol{e})^{\mathrm{T}}]$，T表示向量或矩阵的转置．根据客观情况，总可假定各个风险证券的收益率之间是线性不相关的，则 $\boldsymbol{V}$ 为正定矩阵．

投资决策就是确定投资各种证券的比例(或权重)．用 ω_i 表示投资在风险证券 i 上的投资权重，并令 $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_N)^{\mathrm{T}}$，则 $1-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{l}$ 为投资到无风险证券上的权重．$\boldsymbol{l}$ 表示每个分量均为1的 N 维向量．如果 $\omega_i<0$，则表示在证券 i 上为“空头”(short position)；如果 $1-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{l}<0$，则表示投资者按利率 r_f 借款．为叙述方便起见，可把投资到风险证券上的权重向量 $\boldsymbol{\omega}$ 称为**一投资组合**．

令 $\boldsymbol{\omega}$ 为一投资组合，则它的资产收益率 $r(\boldsymbol{\omega})$(简称 $\boldsymbol{\omega}$ 的收益率)及收益率的方差 $\sigma^2(\boldsymbol{\omega})$ 分别为

$$r(\boldsymbol{\omega})=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{r}+(1-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{l})r_f,\qquad \sigma^2(\boldsymbol{\omega})=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\omega}.\tag{1-1}$$

如果用 $\mu(\boldsymbol{\omega})$ 表示 $\boldsymbol{\omega}$ 的预期收益率，即 $\mu(\boldsymbol{\omega})=E[r(\boldsymbol{\omega})]$，则称 $\mu(\boldsymbol{\omega})-r_f$ 为 $\boldsymbol{\omega}$ 的**预期超额收益率**，称它与 $\sigma(\boldsymbol{\omega})$ 之比为 $\boldsymbol{\omega}$ 的夏普比率(Sharpe ratio)．

1.1.2 均值-方差分析

令 $\boldsymbol{\omega}$ 为一投资组合，如果在所有与 $\boldsymbol{\omega}$ 有相同的预期收益的投资组合中，$\boldsymbol{\omega}$ 的收益率的方差最小，则称 $\boldsymbol{\omega}$ 为**最小方差投资组合**．

下面分两种情形研究最小方差投资组合．

(1)只在风险证券上的投资情形．这时预期收益率为 μ 的最小方差投资组合 $\boldsymbol{\omega}(\mu)$ 是如下带线性约束条件的二次规划问题的解：

$$\begin{cases}\min_{\boldsymbol{\omega}}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\omega};\\ \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{e}=\mu,\qquad \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{l}=1.\end{cases}\tag{1-2}$$

用拉格朗日(J.L.Lagrange)乘子法求得(1-2)式的解为

$$\boldsymbol{\omega}(\mu)=\boldsymbol{g}+\mu\boldsymbol{h},\tag{1-3}$$

其中

$$\boldsymbol{g}=\frac{1}{D}(B\boldsymbol{V}^{-1}\boldsymbol{l}-A\boldsymbol{V}^{-1}\boldsymbol{e}),\quad \boldsymbol{h}=\frac{1}{D}(C\boldsymbol{V}^{-1}\boldsymbol{e}-A\boldsymbol{V}^{-1}\boldsymbol{l}),$$

$$A=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{e}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{l},\quad B=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{e},$$

$$C=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{l},\quad D=BC-A^2.$$

易知 $B>0, C>0, D>0$.

由(1-3)式可立刻推得如下的**两基金分离定理**(two-fund separation theorem).

定理1 设 $N\geqslant 2$, $\boldsymbol{\omega}^{(1)}$和 $\boldsymbol{\omega}^{(2)}$为两个不同的最小方差投资组合,则任一最小方差投资组合 $\boldsymbol{\omega}$ 可由$\boldsymbol{\omega}^{(1)}$和 $\boldsymbol{\omega}^{(2)}$的线性组合表示,即存在实数 α,使得 $\boldsymbol{\omega}=\alpha\boldsymbol{\omega}^{(1)}+(1-\alpha)\boldsymbol{\omega}^{(2)}$,其中 α 为如下方程的唯一解:

$$\mu(\boldsymbol{\omega})=\alpha\mu(\boldsymbol{\omega}^{(1)})+(1-\alpha)\mu(\boldsymbol{\omega}^{(2)}).$$

反之,对任何实数 α, $\alpha\boldsymbol{\omega}^{(1)}+(1-\alpha)\boldsymbol{\omega}^{(2)}$为一最小方差投资组合.

设 p,q 为两个最小方差投资组合,则由(1-3)式知它们的收益率 $r(p)$及 $r(q)$之间的协方差为

$$\begin{aligned}\mathrm{cov}(r(p),r(q))&=\boldsymbol{\omega}(p)^{\mathrm{T}}V\boldsymbol{\omega}(q)\\&=\frac{C}{D}\{E[r(p)]-A/C\}\{E[r(q)]-A/C\}+\frac{1}{C}.\end{aligned}\tag{1-4}$$

特别有

$$\frac{\sigma^2(\mu)}{1/C}-\frac{(\mu-A/C)^2}{D/C^2}=1,\tag{1-5}$$

其中,$\sigma^2(\mu)$表示预期收益率为 μ 的最小方差投资组合的收益率的方差.由(1-5)式看出,$(\sigma(\mu),\mu)$构成 σ-μ 平面上的一条双曲线(见图1-1中的虚线),称之为**投资组合边界**.与总体上方差最小的投资组合对应的点为$(\sqrt{1/C},A/C)$.由于总假定投资者在相同风险前提下偏好于预期收益率高的投资组合,可只需考虑那些预期收益率大于或等于 A/C 的最小方差投资组合,称它们为**有效投资组合**(effecient portfolio),它对应于投资组合边界的上半部分.显然,有效投资组合同时满足两个条件:在相同风险前提下,预期收益率最高;在相同预期收益率水平下,风险最小.

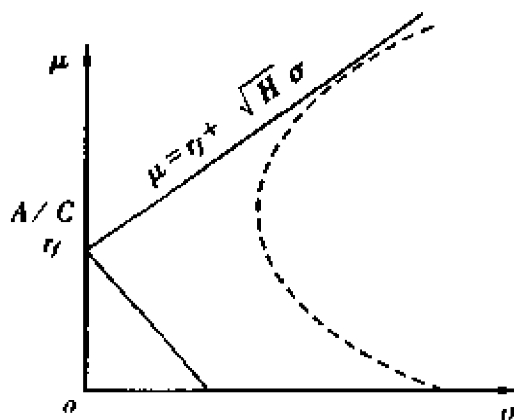


图1-1 投资组合边界与资本市场线

(2) 可同时在无风险证券上投资的情形.这时总可假定风险投资的预期收益率严格大于或等于无风险证券的收益率,即假定 $r_f<A/C$.这时预期收益率为 μ 的最小方差投资组合 $\boldsymbol{\omega}(\mu)$是如下带线性约束条件的二次规划问题的解:

$$\begin{cases}\min_{\boldsymbol{\omega}}\boldsymbol{\omega}^{\mathrm{T}}V\boldsymbol{\omega};\\ \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{e}+(1-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{l})r_f=\mu.\end{cases}\tag{1-6}$$

用拉格朗日乘子法求得(1-6)式的解为

$$\boldsymbol{\omega}(\mu)=V^{-1}(\boldsymbol{e}-r_f\boldsymbol{l})\frac{(\mu-r_f)}{H},\tag{1-7}$$

其中

$$H=(\boldsymbol{e}-r_f\boldsymbol{l})^{\mathrm{T}}V^{-1}(\boldsymbol{e}-r_f\boldsymbol{l})=B-2r_fA+r_f^2C.\tag{1-8}$$

由(1-7)式得

$$\sigma^2(\mu)=\boldsymbol{\omega}(\mu)^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\omega}(\mu)=\frac{(\mu-r_f)^2}{H},$$

从而有

$$\sigma(\mu)=\frac{1}{\sqrt{H}}|\mu-r_f|. \tag{1-9}$$

由(1-9)式看出,$(\sigma(\mu),\mu)$构成平面上从(O,r_f)出发斜率分别为$\sqrt{H}$及$-\sqrt{H}$的两条射线(见图1-1). 只需考虑那些预期收益率大于或等于r_f的最小方差投资组合,称为有效投资组合,与它们相应的点$(\sigma(\mu),\mu)$位于斜率为$\sqrt{H}$的那条射线上. 这条射线称为**资本市场线**(capital market line),它的斜率$\sqrt{H}$是有效投资组合的预期超额收益率与风险的共同比率,称为**风险的市场价格**(market price of risk). 与之相对照的是,无风险证券的收益率可以看成是时间的价格. 由(1-9)式知,有效投资组合$\boldsymbol{\omega}$的预期收益率$\mu(\boldsymbol{\omega})$在扣除**风险报酬**(risk premium)$\sigma(\boldsymbol{\omega})\sqrt{H}$后应等于无风险证券的收益率(时间的价格). 资本市场线与投资组合边界相切于点$(\sqrt{H}/(A-r_fC),(B-r_fA)/(A-r_fC))$上,与该点对应的有效投资组合称为**切点投资组合**(tangency portfolio),记为$\boldsymbol{\omega}^*$. 有

$$\boldsymbol{\omega}^*=\frac{1}{(A-r_fC)}\boldsymbol{V}^{-1}(\boldsymbol{e}-r_f\boldsymbol{I}). \tag{1-10}$$

由(1-7)式知$\boldsymbol{\omega}(\mu)$关于μ是线性的,故在现在情形下也有与定理1相类似的两基金分离定理. 特别地,任一有效投资组合$\boldsymbol{\omega}$可由无风险证券和切点投资组合生成. 此外,当且仅当$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{I}>1$时,$(\sigma(\boldsymbol{\omega}),\mu(\boldsymbol{\omega}))$位于切点的上方. 这时,投资者要按利率$r_f$借款,借款额占总投资额的比例为$\boldsymbol{\omega}'\boldsymbol{I}-1$.

1.1.3 最优投资组合

投资者如何选择有效投资组合取决于他的风险偏好,经济学中通常用所谓的期望效用函数描述. 投资者的目标是使他的收益期望效用最大化. 如果投资者是风险回避者(risk averter),则他的效用函数为一严格增的凹函数. 假定各种证券的收益率服从联合正态分布,或投资者的效用函数为一个二次函数,则期望效用函数化为收益率的均值和标准差的二元函数$u(\sigma,\mu)$,称之为**均值-方差效用函数**. 这时使函数u最大化归结为在预期收益率与风险之间进行权衡. 通常有$\partial u/\partial\sigma<0$, $\partial u/\partial\mu>0$,于是对每个常数c,方程$u(\sigma,\mu)=c$确定了σ-μ平面上的一条有正斜率的曲线,称为无差异曲线. 无差异曲线上的不同点对应于不同的投资组合,但它们的期望效用相同,即投资者对它们有相同的满意度. 容易证明,存在唯一的一条无差异曲线与资本市场线相切,该切点(σ,μ)为如下方程的唯一解:

$$-\frac{\partial u/\partial\sigma}{\partial u/\partial\mu}=\sqrt{H}=\frac{\mu-r_f}{\sigma}. \tag{1-11}$$

对此,μ由(1-7)式求得的有效投资组合即为**最优投资组合**(见图1-2).

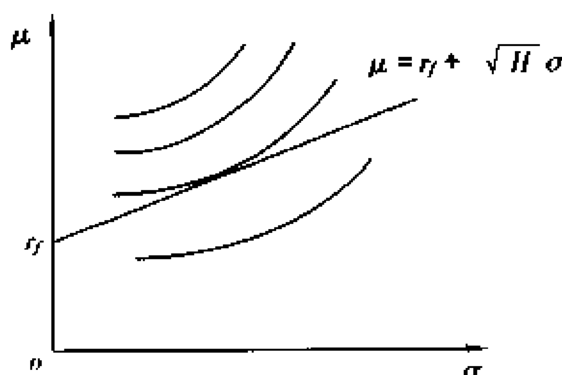


图 1-2 无差异曲线与最优投资组合

(1-11)式表明，对最优投资组合来说，风险与预期收益率之间的边际替代率必须等于风险的市场价格．例如设投资者的均值-方差效用函数为

$$u(\sigma,\mu)=\mu-\frac{1}{\tau}\sigma^2, \tag{1-12}$$

这里 $\tau>0$ 称为投资者的**风险容忍度**（τ^{-1}称为**风险厌恶系数**），则方程(1-11)的解为

$$\mu=r_f+\frac{\tau}{2}H,\quad \sigma=\frac{\tau}{2}\sqrt{H}. \tag{1-13}$$

1.2 资本资产定价模型

1.2.1 市场投资组合

设市场上风险证券 i 的当前总价值为 W_i，无风险证券的当前总价值为 W_f，令

$$w_i = W_i\Big/\left(\sum_{j=1}^{N}W_i + W_f\right), \tag{1-14}$$

则投资组合 $\boldsymbol{w}=(w_1,\cdots,w_N)$称为**市场投资组合**．在实际应用中，市场投资组合可用某一指数基金来近似，因为指数基金是一种拥有广泛代表性的风险证券的共同基金(mutual fund)．下面将要证明，在**竞争均衡**(competitive equilibrium)市场中，市场投资组合是有效的．为此假定市场无磨擦(即无交易费，无税金，证券可以任意分割，对借款和卖空不加限制，借贷利率相同)，投资者对市场中每个证券的收益率有相同的预期且都采用使各自的均值-方差效用函数达到最大的最优投资组合(从而为有效投资组合)．如果这时市场中证券数量的总供求相等(即市场结清)，则称市场是竞争均衡的．设市场中有 K 个投资者，各自的投资额为 $Z^{(1)},\cdots,Z^{(K)}$，各自的最优投资组合为 $\boldsymbol{\omega}^{(1)},\cdots,\boldsymbol{\omega}^{(K)}$．于是有

$$W_i=\sum_{k=1}^{K}\omega_i^{(k)}Z^{(k)},\quad \sum_{k=1}^{K}Z^{(k)}=\sum_{j=1}^{N}W_j+W_f.$$

令 $\alpha_k = Z^{(k)} \Big/ \sum_{k=1}^{K} Z^{(k)}$，则由(1-14)式得

$$w_i = \sum_{k=1}^{K} \omega_i^{(k)} Z^{(k)} \Big/ \sum_{k=1}^{K} Z^{(k)} = \sum_{k=1}^{K} \alpha_k \omega_i^{(k)}. \tag{1-15}$$

因此，作为 K 个有效投资组合的一个凸组合，市场投资组合是有效的．如果无风险证券净供给为零(即市场上按利率 r_f 借贷总额相抵)，市场投资组合就是切点投资组合(见图 1-1)．

下面简称市场投资组合的风险为**市场风险**，称市场投资组合的预期超额收益率为市场风险报酬，后者就等于市场风险与风险的市场价格的乘积．

由两基金分离定理，用无风险证券和市场投资组合，可以产生任意有效投资组合．例如，设每个投资者的均值-方差效用函数由形如(1-12)式给出，风险容忍度分别为 $\tau_1, \cdots, \tau_k$，于是由(1-13)式及(1-7)式得

$$\omega^{(k)} = \frac{\tau_k}{2} V^{-1}(e - r_f l) = \frac{\tau_k H}{2(\mu(w) - r_f)} w. \tag{1-16}$$

这时由(1-15)式及(1-16)式有

$$\omega = \frac{\tau H}{2(\mu(w) - r_f)} w, \tag{1-17}$$

其中

$$\tau = \sum_{k=1}^{K} \alpha_k \tau_k. \tag{1-18}$$

对比(1-17)式和(1-16)式可看出，市场投资组合可以看成风险容忍度为 τ 的投资者的最优投资组合，其中 τ 为市场中的投资者的风险容忍度按其投资额占总投资额的比例的加权平均．

1.2.2 风险资产的 β 系数

设 ω 为任一投资组合(不必为最小方差投资组合)，ω' 为一最小方差投资组合．由(1-7)式得

$$\begin{aligned}\operatorname{cov}(r(\omega), r(\omega')) &= \omega^{\mathrm{T}} V \omega' = \omega^{\mathrm{T}}(e - r_f l)\frac{\mu(\omega') - r_f}{H} \\ &= \frac{(\mu(\omega) - r_f)(\mu(\omega') - r_f)}{H}.\end{aligned} \tag{1-19}$$

于是，当 $\mu(\omega) \neq r_f$ 时，有

$$\mu(\omega) = r_f + \beta_{\omega,\omega'}(\mu(\omega') - r_f), \tag{1-20}$$

其中

$$\beta_{\omega,\omega'} = \frac{\operatorname{cov}(r(\omega), r(\omega'))}{\sigma^2(\omega')}, \tag{1-21}$$

特别地，令 ω' 为市场投资组合 w，则有

$$\mu(\omega) = r_f + \beta_{\omega,w}(\mu(\omega) - r_f), \tag{1-22}$$

称 $\beta_{\omega,w}$ 为投资组合 ω 的 **β 系数**，它等于 ω 的预期超额收益率 $\mu(\omega) - r_f$ 与市场风

险报酬$\mu(w)-r_f$之比．令$\rho_{r(\omega),r(w)}$表示$r(\omega)$与$r(w)$的相关系数，即

$$\rho_{r(\omega),r(w)}=\frac{\operatorname{cov}(r(\omega),r(w))}{\sigma(\omega)\sigma(w)}=\beta_{\omega,w}\frac{\sigma(w)}{\sigma(\omega)},\tag{1-23}$$

则由(1-9)式知，可把(1-22)式改写成

$$\mu(\omega)=r_f+\rho_{r(\omega),r(w)}\sigma(\omega)\sqrt{H}=r_f+\beta_{\omega,w}\sigma(w)\sqrt{H}.\tag{1-24}$$

称$\rho_{r(\omega),r(w)}\sigma(w)$为投资组合$\omega$的**系统风险**，称$\sqrt{1-\rho^2_{r(\omega),r(w)}}\sigma(\omega)$为$\omega$的**非系统风险**．由(1-23)式知，$\omega$的$\beta$系数也等于$\omega$的系统风险与市场风险之比．由(1-9)式知，当且仅当投资组合是有效时，非系统风险才为零．此外由(1-24)式知，投资组合的预期超额收益率$(\mu(\omega)-r_f)$只取决于投资组合的系统风险，而不取决于它的总风险，后者还含有投资组合的非系统风险．通常把关系式(1-22)或式(1-24)称为**资本资产定价模型(CAPM)**．如果用β系数作为坐标平面的横轴，预期收益率μ作为纵轴，则截距为r_f、斜率为$\sigma(w)\sqrt{H}$(市场的风险报酬)的直线称为**证券市场线**(security market line)(见图1-3)．

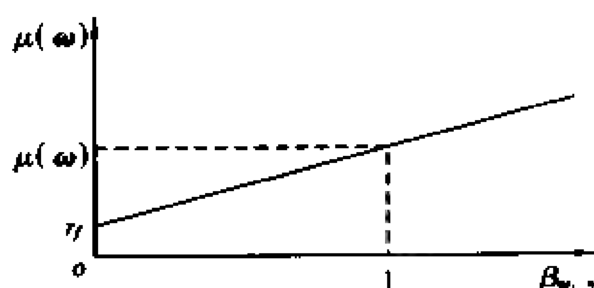


图1-3 证券市场线

特别地，令$\beta_i=\operatorname{cov}(r_i,r(w))/\sigma^2(w)$，称$\beta_i$为证券$i$的**$\beta$系数**．由(1-23)式得

$$e_i=r_f+\beta_i(\mu(w)-r_f).\tag{1-25}$$

$\beta_i\sigma(w)$为证券i的系统风险，它是由整个市场大环境引起的，不能通过分散投资来消除，而证券i的非系统风险$\sqrt{\sigma^2(r_i)-\beta_i^2\sigma^2(w)}$是由公司的内在因素造成的，它能被投资者通过分散投资来消除．证券的β系数可以作为证券的相对系统风险的一个度量．容易看出，投资组合ω的β系数，实际上等于证券的β系数关于证券的投资权重的加权平均(注意无风险证券的β系数为零)，即有

$$\beta_{\omega,w}=\sum_{i=1}^{N}\omega_i\beta_i.$$

证券的β系数可以由公司和证券市场的历史数据用统计方法作出估计，市场风险报酬$(\mu(w)-r_f)$可以用某一指数基金的超额预期收益率作近似．

1.2.3 CAPM的应用

CAPM的首要应用是对风险资产定价．假定市场是竞争均衡的，设X为一风险资产在单个期间末的价格，它是一随机变量．假定它在市场中的β系数已知，希望确定X的当前价格X_0．由(1-24)式知，X的收益率$r=X/X_0-1$必须满足如下市

场均衡条件：

$$E[r] = r_f + \rho_{r,r(w)}\sigma(r)\sqrt{H}, \tag{1-26}$$

称 X_0 为 X 的当前均衡价格(或现值). 由于

$$E[X] = (1 + E[r])X_0, \quad \rho_{r,r(w)} = \rho_{X,r(w)}, \quad \sigma(X) = X_0\sigma(r),$$

故由(1-26)式得

$$X_0 = \frac{E[X]}{1 + r_f + \rho_{X,r(w)}\sigma(w)\sqrt{H}}. \tag{1-27}$$

均衡定价对公司做资本预算很重要,因为公司的财务目标是资产价值最大化. 公司应选择那些项目进行投资,它们产生的未来不确定资产的当前均衡价格高于成本预算 .

CAPM 的另一应用是投资者可利用均衡价格与实际价格的对比,对证券或风险资产(如共同基金)进行评估,发现被低估或高估的,从卖高买低中获益 . 但需注意的是,不能只投资于个别被认为低估了的证券,因为单个证券的非系统风险不容忽视 .

1.3 套利定价理论

前面介绍的 CAPM 是一个单因子模型,它实际隐含地假定了影响证券收益率的共同因素是单个市场因素,而且这一因素实际上是不可观测的 . 但事实上有多种共同因素(如国民生产总值、就业率、银行利率及通货膨胀指数)都会影响市场上大多数证券的收益率 . 这决定了 CAPM 在应用上的局限性 .CAPM 的更大缺陷是它理想化地假定了投资者都是风险厌恶者且按均值-方差效用函数最大化准则决定最优投资组合策略,还假定了市场是竞争均衡的 . 罗斯(S. Ross)完全放弃这些不切实际的假定,于 1976 年对证券市场中的证券收益率提出了下述多因子模型：

$$r_i = e_i + \sum_{j=1}^{M} b_{ij}f_j + \varepsilon_i, \quad 1 \leqslant i \leqslant N, \tag{1-28}$$

其中,e_i 为证券 i 的预期收益率;$f_1,\cdots,f_M$ 表示影响证券收益率的共同因素,每个 f_j 为零均值随机变量;b_{ij}为证券 i 对因素 j 的敏感系数;$\varepsilon_1,\cdots,\varepsilon_N$ 为互不相关的零均值随机变量,代表模型误差,且与$(f_1,\cdots,f_M)$也不相关 . 基于这一模型,罗斯提出了套利定价理论(APT).APT 的基本结果可以粗略地描述如下:假定市场上证券数量 N 相对因素个数 M 来说非常大,且市场是渐进无套利的(即当 N 趋于无穷大时,套利机会逐渐消失),则存在 M 个投资组合 $\omega(1),\cdots,\omega(M)$,使得第 j 个投资组合对第 j 个因素的敏感系数为 1,对其他因素的敏感系数为零,且绝大多数证券的预期收益率与证券对市场共同因素的敏感度之间近似地存在如下线性关系：

$$e_i = r_f + \sum_{j=1}^{M} b_{ij}(\lambda_j - r_f), \tag{1-29}$$

其中,λ_j 为 $\omega(j)$的预期收益率,$\lambda_j - r_f$ 代表第 j 个因素的风险报酬 . 与 CAPM 模型类似,(1-29)式可用于风险资产的定价和评估 .

从数学上严格叙述 APT 需要假定市场中有无穷多个证券,每个证券的收益率满足方程(1-28),且假定 $E\varepsilon_i^2$ 一致有界 . 如果市场不存在渐进套利机会,则 APT 理论断言:存在与 n 无关的 $M+1$ 个常数 $\mu_0,\cdots,\mu_M$,使得

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(e_i-\mu_0-\sum_{j=1}^{M}b_{ij}\mu_j)^2=0. \tag{1-30}$$

如果市场上存在无风险证券,则 μ_0 为无风险证券的收益率 .

2 期权定价理论

本章首先通过二叉树模型和离散时间模型介绍金融经济学的基本概念和期权的风险中性定价原理;其次介绍布莱克-索尔斯模型和欧式期权定价公式及其实际应用;然后介绍未定权益定价和复制的鞅方法,介绍美式期权的定价,并给出若干特异期权的定价公式;最后介绍利率的期限结构和利率衍生资产的定价 .

2.1 离散时间模型

2.1.1 二叉树模型

二叉树模型(binomial-tree model)是一个最简单和直观的证券市场的数学模型 . 该模型是 1979 年由考克斯(J. Cox)、罗斯(S. Ross)和鲁宾斯坦(M. Rubinstein)引进的,他们用这一模型给出了布莱克-索尔斯的期权定价公式的一个简单和初等的推导 .

设市场中只有一种风险证券(例如股票)和一种无风险证券(如政府债券),后者在每个期间的利率为一固定常数 r . 用 S_0 表示风险证券在当前时刻 0 的价格, S_n 表示风险证券在时刻 n(即第 n 个期间末)的价格 . 假定在每个期间,股票价格变动只有两种可能,且相对幅度不随期间改变,即存在正数 $d,u(d<u)$,使得对每个 $n\geqslant 0$,有 $S_{n+1}=uS_n$ 或 $S_{n+1}=dS_n$(见图 2-1). 假定这两种情形的概率分别为 p 和 $1-p(p>0)$. 为了市场中无套利机会,必须有 $d<1+r<u$.

考虑一金融合约,它在时刻 1(第 1 个期间末)的价值依赖于股票在时刻 1 的价格:当股票价格为 uS_0 时,它为 ξ_u;当股票价格为 dS_0 时,它为 ξ_d. 要研究的问题是:如何合理确定合约的当前价格,使得市场中仍无套利机会 . 为此,构造一投资组合,它由卖空合约和买进 α_0 份股票构成,使得它在时刻 1 的资产是无风险的(即非随机的). 显然,α_0 应由如下方程确定:

$$\alpha_0 uS_0-\xi_u=\alpha_0 dS_0-\xi_d,$$

其解为 $\alpha_0=(\xi_u-\xi_d)/((u-d)S_0)$. 于是,该投资组合在时刻 1 的资产 X_1 为$(d\xi_u-u\xi_d)/(u-d)$. 为了确保市场无套利,该投资组合在时刻 0 的资产 X_0 应为 $X_1/(1+r)$,因为若将 X_0 投资到无风险证券上,在时刻 1 也应获得 X_1. 由此推得合约的当

前价格 C_0 为

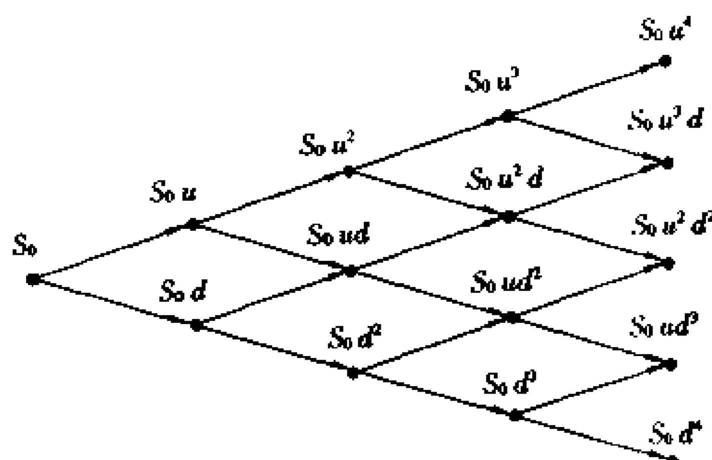


图 2-1 二叉树模型

$$C_0 = \alpha_0 S_0 - X_0 = \frac{(1+r-d)\xi_u + [u-(1+r)]\xi_d}{(1+r)(u-d)}. \tag{2-1}$$

这一定价方法称为**套利定价**(arbitrage pricing). 从定价公式(2-1)可看出,该价格不依赖于概率 p 的大小,即不依赖于投资者对市场中证券收益率的预期.

令 $q=(1+r-d)/(u-d)$,则(2-1)式可改写成

$$C_0 = (1+r)^{-1}(q\xi_u + (1-q)\xi_d). \tag{2-2}$$

如果将 q 和 $1-q$ 看成为股票价格上下变动的概率(它们构成状态空间上一概率测度,记为 P^*),则(2-2)式表明:合约的当前价格 C_0 为合约的**贴现价值**(discount value)在概率测度 P^* 下的数学期望. 由于 $qu+(1-q)d=1+r$,易知 P^* 是唯一的概率测度,使得 $E^*[(1+r)^{-1}S_1]=S_0$.

下面考虑多期间(multiperiod)交易情形. 设 ξ 为在时刻 N 到期的一合约的价值,令 Ω 表示到 N 时刻以前的股票价格各种(2^N 种)可能涨跌走势,则原来的概率分布$(p,1-p)$和新的概率分布$(q,1-q)$分别决定了 Ω 上的一概率测度 P 和 P^*. 容易证明,P^* 是 Ω 上唯一的概率测度,使得股票的贴现价格序列$\{(1+r)^{-n}(S_n), n\leq N\}$在 P^* 下为一鞅,且合约在时刻 n 的无套利价格为

$$C_n = (1+r)^{-(N-n)}E^*[\xi|\mathscr{F}_n]. \tag{2-3}$$

称 P^* 为**风险中性概率测度**(risk-neutral probability measure)或**鞅测度**(martingale measure). 在 P^* 下,合约的贴现价格序列也为鞅.(2-3)式是所谓的**风险中性定价原理**(risk-neutral valuation principle)的一个例子. 若 $\xi=f(S_N)$,则容易由(2-3)式推得

$$C_n = (1+r)^{-(N-n)}\sum_{j=0}^{N-n}\binom{N-n}{j}q^j(1-q)^{N-n-j}f(S_n u^j d^{N-n-j}). \tag{2-4}$$

2.1.2 一般的离散时间模型

考虑在 N 期间交易的证券市场. 直到时刻 N 的市场不确定性由一概率空间$(\Omega,\mathscr{F},P)$表示,其中 Ω 表示所有可能状态的集合. 令 $\mathscr{F}_n$ 为 $\mathscr{F}$ 的一子 σ 代数,它代

表直到时刻 n 的市场信息，则 $\{\mathscr{F}_n, 0\leqslant n\leqslant N\}$ 构成 Ω 上的一非降 σ 代数流(filtration). 为方便起见，令 $\mathscr{F}_{-1}=\mathscr{F}_0$.

设市场上有 $d+1$ 种证券，它们在时刻 n 的价格构成一 R^{d+1} 值非负随机向量 $S_n=(S_n^0,\cdots,S_n^d)^{\mathrm{T}}$. 证券 0 为一无风险证券(如债券)，它在时刻 n 的价格约定为 $S_n^0=(1+r)^n$，记为 β_n，其中 $r>0$ 为它在单个期间的收益率. 用 γ_n 表示贴现因子 $(1+r)^{-n}$，其余证券为风险证券.

一个交易策略是一列投资组合，它是一个可料的(predictable) $d+1$ 维向量序列

$$\boldsymbol{\phi}=\{(\phi_n^0,\cdots,\phi_n^d)^{\mathrm{T}}, 0\leqslant n\leqslant N\}.$$

即每个 ϕ_n^i 为 $\mathscr{F}_{n-1}$ 可测，它表示在时刻 n 的投资组合中拥有证券 i 的份数. 对投资组合序列作可料性假定表明，在时刻 n 作决策时只能利用限制在时刻 n 之前(即直到时刻 $n-1$)的市场信息. 若 $\phi_n^0<0$，则表明卖空 $|\phi_n^0|$ 份债券；若 $i\geqslant 1, \phi_n^i<0$，则表明卖空 $|\phi_n^i|$ 份证券 i. 在时刻 n 投资组合的财富为

$$V_n(\boldsymbol{\phi}) = \boldsymbol{\phi}_n^{\mathrm{T}}S_n = \sum_{i=0}^{d}\phi_n^i S_n^i, \tag{2-5}$$

其贴现值为

$$\widetilde{V}_n(\boldsymbol{\phi})=\gamma_n V_n(\boldsymbol{\phi})=\boldsymbol{\phi}_n\cdot\widetilde{S}_n, \tag{2-6}$$

其中，$\widetilde{S}_n=(1,\gamma_n S_n^1,\cdots,\gamma_n S_n^d)^{\mathrm{T}}$.

一交易策略 $\boldsymbol{\phi}$ 称为**自融资的**(self-financing)，如果

$$\boldsymbol{\phi}_n^{\mathrm{T}}S_n=\boldsymbol{\phi}_{n+1}^{\mathrm{T}}S_n, \quad \forall 0\leqslant n\leqslant N-1, \tag{2-7}$$

即投资者每次调整投资组合时，既不追加投资又不抽走资金. (2-7)式等价于

$$V_n(\boldsymbol{\phi}) = V_0(\boldsymbol{\phi}) + \sum_{i=1}^{n}\boldsymbol{\phi}_i^{\mathrm{T}}\Delta S_i, \quad \forall 1 \leqslant n \leqslant N, \tag{2-8}$$

或

$$\widetilde{V}_n(\boldsymbol{\phi}) = V_0(\boldsymbol{\phi}) + \sum_{i=1}^{n}\boldsymbol{\phi}_i\cdot\Delta\widetilde{S}_i, \quad \forall 1 \leqslant n \leqslant N, \tag{2-9}$$

其中，$\Delta S_i=S_i-S_{i-1}$，$\Delta\widetilde{S}_i=\widetilde{S}_i-\widetilde{S}_{i-1}$.

一交易策略 $\boldsymbol{\phi}$ 称为容许的(admissible)，如果其财富过程非负. 初始财富为零、终了时刻财富非零的自融资容许策略称为套利策略.

定理 1 当且仅当存在一个与 P 等价的概率测度 P^* 使得 $(\widetilde{S}_n)_{0\leqslant n\leqslant N}$ 为一个 P^* 鞅时，市场无套利. 这时可选取 P^*，使得它关于 P 的拉东-尼古丁(Radon-Nikodym)导数 $\mathrm{d}P^*/\mathrm{d}P$ 有界.

定理 1 称为**资产定价基本定理**(fundamental theorem of asset pricing)，它给出了无套利市场的鞅刻画.

称定理 1 中的概率测度 P^* 为市场的**等价鞅测度**. 一般说来，等价鞅测度不唯一.

下面假定市场无套利，即存在市场的等价鞅测度. 执行时刻为 N 的一欧式未

定权益(European contingent claim)是一 $\mathscr{F}_N$ 可测的非负随机变量,它表示在未来时刻 N 可实现的权益．如果未定权益的价值依赖于一个或几个标的资产(underlying asset),则也称此未定权益为衍生资产(derivative asset).衍生资产的一个典型例子是**期权**(option),它分为**买权**(call option)和**卖权**(put option).期权是基于某一标的资产(如股票)的一金融合约．买权的持有者有权(但无义务)在合约到期日(expiration date 或 maturity) N 从合约卖方按**约定价**(striking price) K 买一份标的资产．因此买权的权益为 $(S_N-K)^+$．只有当标的资产在合约到期时价格高于约定价时买方才会执行合约．类似地,卖权的权益为 $(K-S_T)^+$,它的持有者有权(但无义务)在合约到期日 T 按约定价 K 卖一份标的资产给合约卖方．如果存在一自融资策略使其在时刻 T 的财富等于 ξ,则未定权益 ξ 称为**可复制的**(replicatable),可以证明,复制一未定权益的自融资策略必为容许的,且在任一等价鞅测度 P^* 下,它的贴现财富过程为一鞅．因此,如果 ξ 为一个可复制的未定权益,则所有复制它的自融资策略的财富序列 (V_n) 是相同的,且有

$$V_n=\beta_n E^*[\gamma_N\ \xi|\mathscr{F}_n],\tag{2-10}$$

其中 $\gamma_N=(1+r)^{-N}$,E^* 为对应于 P^* 的期望算子,P^* 是任一等价鞅测度．这时,为了保持市场无套利,ξ 在时刻 n 的价格必须定义为复制策略在时刻 n 的财富 V_n.称(2-10)式为风险中性定价公式．

令 $\mathscr{M}$ 表示市场的等价鞅测度全体．对不可复制的未定权益 ξ,令

$$V_n^s=\operatorname*{ess.sup}_{Q\in\mathscr{M}}\beta_n E_Q[\gamma_N\xi|\mathscr{F}_n],$$

$$V_n^b=\operatorname*{ess.inf}_{Q\in\mathscr{M}}\beta_n E_Q[\gamma_N\xi|\mathscr{F}_n],$$

分别称 V_n^s,V_n^b 为 ξ 在时刻 n 的卖方价和买方价．

设市场无套利．如果每个未定权益 ξ 都是可复制的,市场称为完备的(complete).可以证明,为了市场是完备的,必须且只需存在唯一的等价鞅测度．

下面讨论**美式未定权益**(American contingent claim)的定价．与欧式未定权益不同的是,美式未定权益在合约到期之前的任何时刻都可执行．一般说来,到期时刻为 N 的美式未定权益可用关于 $(\mathscr{F}_n)$ 适应的非负随机变量序列 (Z_n) 描述,Z_n 表示在时刻 n 执行合约所获得的权益,即合约的卖方向买方付给 Z_n.例如,对于股票的美式买权,$Z_n=(S_n-K)^+$,其中 S_n 为股票在时刻 n 的价格,K 为期权合约的约定价格或**执行价格**(exercise price).可用倒向归纳法来对美式未定权益定价．假定市场是完备的,P^* 为唯一的等价鞅测度．令 U_n 表示美式未定权益在时刻 n 的卖方价格,则 $U_N=Z_N$.如果合约卖方要确保他能在时刻 $N-1$ 支付 Z_{N-1} 和在时刻 N 支付 Z_N,则由(2-10)式,应定义

$$U_{N-1}=\max(Z_{N-1},\beta_{N-1}E^*[\tilde{Z}_N|\mathscr{F}_{N-1}]).\tag{2-11}$$

由归纳法得

$$U_n=\max(Z_n,\beta_n E^*[\gamma_{n+1}U_{n+1}|\mathscr{F}_n]).\tag{2-12}$$

定理 2 美式未定权益的贴现价格序列 $(\tilde{U}_n)_{0\leqslant n\leqslant N}$ 为 P^* 上鞅．它是从上控制

序列$(\tilde{Z}_n)_{0\leqslant n\leqslant N}$的最小 P^* 上鞅．

2.2 布莱克-索尔斯模型

2.2.1 布莱克-索尔斯方程和定价公式

现在考虑连续时间市场模型．设市场上只有两种证券：一是风险证券（如股票），二是无风险证券（如债券）．假定债券的初始价格为 1，连续复利率为 r．假定股票不分红，股票价格过程 S_t 满足如下的伊藤（Itô，K.）随机微分方程：

$$dS_t = S_t(\mu dt + \sigma dB_t), \tag{2-13}$$

其中 $S_0>0$，μ 和 σ 为常数，(B_t)为定义在带 σ 代数流的概率空间$(\Omega,\mathscr{F},(\mathscr{F}_t),P)$上的一标准布朗运动．过程$(S_t)$称为**几何布朗运动**．有

$$S_t = S_0\exp\left\{\left(u-\frac{\sigma^2}{2}\right)t+\sigma B_t\right\}, \tag{2-14}$$

从而 $\log(S_t)$服从正态分布．称 μ 为股票的（瞬时）预期收益率，σ 为股票的波幅（volatility）．注意股票的预期连续复利率是 $E\log\frac{S_t}{S_0}$（即 $\mu-\frac{\sigma^2}{2}$），它不同于股票的瞬时预期收益率 μ．令 $\beta_t=e^{rt}$，它表示债券在时刻 t 的价格．

假定市场无磨擦（frictionless），即无交易费，无税金，无买卖价差（bid-ask spread），允许卖空，证券可以任意分割．此外，假定可连续交易．一个交易策略为一对 $\mathscr{F}_t$ 适应过程$\{a,b\}$，满足 $a\in\mathscr{L}^2[0,T]$，$b\in\mathscr{L}^1[0,T]$．其中 $a(t)$和 $b(t)$分别表示在时刻 t 的投资组合中拥有股票和债券的份额．投资组合$\{a(t),b(t)\}$的财富 V_t 为

$$V_t=a(t)S_t+b(t)\beta_t.$$

如果对一切 t，有

$$dV_t=a(t)dS_t+b(t)d\beta_t,$$

则交易策略称为自融资的．如果对一切 t，$V_t\geqslant 0$，则交易策略称为容许的．

考虑到期时刻为 T 的形如 $\xi=f(S_T)$的一未定权益，其中 $f:R_+\to R_+$ 为一连续函数．布莱克和索尔斯利用伊藤公式导出了一个复制 ξ 的自融资交易策略的财富过程应满足的方程，并将该财富过程定义为 ξ 的价格过程．因为如不这样定义的话，市场将存在套利机会．布莱克和索尔斯证明 ξ 的价格过程可表成$F(t,S_t)$，其中 F 为如下方程的解：

$$F_t(t,x)+rxF_x(t,x)+\frac{1}{2}\sigma^2x^2F_{xx}(t,x)-rF(t,x)=0,$$
$$(t,x)\in[0,T)\times(0,\infty), \tag{2-15}$$

终端条件为

$$F(T,x)=f(x),\quad x\in(0,\infty).$$

方程(2-15)称为**布莱克-索尔斯方程**．特别地，若未定权益 ξ 是欧式买权，即 $\xi=$

$(S_T-K)^+$，其中 K 为期权的执行价格，则它在时刻 t 的价格 $C_t=C(t,S_t)$ 由如下著名的**布莱克-索尔斯公式**给出：

$$C(t,x)=xN(d_1)-Ke^{-r(T-t)}N(d_2), \tag{2-16}$$

其中 $N(z)$ 为标准正态分布函数，

$$\begin{cases} d_1=\dfrac{\log(x/K)+(r+\frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}; \\ d_2=\dfrac{\log(x/K)+(r-\frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}. \end{cases} \tag{2-17}$$

由于买权和卖权的价格满足**卖权-买权平价关系**(put-call parity)

$$S_t+P_t-C_t=Ke^{-r(T-t)},$$

故由(2-17)式得卖权的价格 $P_t=P(t,S_t)$，其中

$$P(t,x)=Ke^{-r(T-t)}N(-d_2)-xN(-d_1). \tag{2-18}$$

一个重要的事实是：股票的预期收益率在布莱克-索尔斯方程及公式中都不出现．默顿进一步发现：若在股票价格模型(2-13)式中将常数 μ 改成一适应过程，布莱克-索尔斯方程及公式保持不变．

下面对布莱克-索尔斯模型稍加推广，即假定 r,μ,σ 不再是常数，而是时间 t 的函数．此外假定股票按复利率 $q(t)$ 连续派息．这时相应的布莱克-索尔斯方程为

$$-r(t)F(t,x)+F_t(t,x)+[(r(t)-q(t)]xF_x(t,x)+\frac{1}{2}\sigma^2(t)x^2F_{xx}(t,x)=0, \tag{2-19}$$

相应的布莱克-索尔斯公式为

$$C(t,x)=\tilde{x}\ N(\tilde{d}_1)-Ke^{-(T-t)\tilde{r}}\tilde{N}(\tilde{d}_2), \tag{2-20}$$

其中

$$\tilde{x}=x\exp\left(-\int_t^T q(s)\mathrm{d}s\right),\quad \tilde{r}=\frac{1}{T-t}\int_t^T r(s)\mathrm{d}s,$$

$\tilde{d}_1$ 和 $\tilde{d}_2$ 的表达式与(2-17)式相同，唯一不同的是(2-17)式中的 x,r,σ^2 现在分别改成了 $\tilde{x}$，$\tilde{r}$ 和 $\frac{1}{T-t}\int_t^T\sigma^2(s)\,\mathrm{d}s$.

2.2.2 布莱克-索尔斯公式的实际应用

与 CAPM 对风险资产定价不同，布莱克-索尔斯公式给出的是期权的无套利价格，它不依赖于投资者对市场中证券收益率的预期，也不依赖于市场的系统风险，风险是中性的．这里，市场均衡不是期权定价的必要条件，而市场无套利却是期权定价的充分条件．因此，布莱克-索尔斯公式具有广泛实用性和可操作性．在布莱克-索尔斯公式中，唯一的未知参数是股票的波幅 σ. 有三种方法对 σ 进行估计：

一是利用股票价格的历史数据用统计方法来估计,得到的结果称为**历史波幅**(historical volatility);二是利用市场上公布的不同到期时刻及不同约定价格的期权价格,通过布莱克-索尔斯公式反解出波幅,再作某种加权平均,这样得到的估计称为**引申波幅**(implied volatility),它反映市场对股票的当前波幅的综合评估,从而可作为对股票未来波幅的一种预测;三是用一种称为 GARCH 模型的统计方法,对股票的波幅进行预报.

布莱克-索尔斯公式的另一主要应用是,它能提供股票价格相对于各种参数变动敏感性的一个度量.这些度量对监控期权的风险暴露(risk exposure)非常有用.在布莱克-索尔斯公式中对各个参数微分,分别得到

$$\Delta = \frac{\partial C}{\partial x} = N(d_1) > 0,$$

$$\Gamma = \frac{\partial^2 C}{\partial x^2} = \frac{1}{x\sigma\sqrt{T-t}} N'(d_1) > 0,$$

$$V = \frac{\partial C}{\partial \sigma} = x\sqrt{T-t}N'(d_1) > 0,$$

$$\rho = \frac{\partial C}{\partial r} = (T-t)e^{-r(T-t)}KN(d_2) > 0,$$

$$\theta = \frac{\partial C}{\partial t} = -\frac{x\sigma}{2\sqrt{T-t}}N'(d_1) - Kre^{-r(T-t)}N(d_2) < 0.$$

其中,Δ 为度量股票价格的单位改变引起期权价格的改变,$N'(x)$为 $N(x)$的导数.V(即 Vega),ρ,θ 分别反映波幅、利率及离期权到期时间长度的变化对期权价格的影响.由伊藤公式易知,在复制或对冲(hedge)买权的交易策略中,时刻 t 的股票持有量为 $\Delta(t,S_t)$.因此,通常把这一对冲交易策略称为 **δ 对冲**.在期权卖方为避免风险而做 δ 对冲时,为减少交易费用,只当 Δ 有较大变动时才对股票持有量作调整.Δ 对股票价格变动的敏感性 Γ 有助于了解调整股票持有量的频繁程度.

期权价格敏感性的另一度量是所谓的弹性 λ,它等于用股票价格的百分比变动除期权价格的百分比变动.由布莱克-索尔斯公式得

$$\lambda = \frac{\partial \log C}{\partial \log x} = \frac{xN(d_1)}{C}.$$

由(2-16)式知,对买权恒有 $\lambda > 1$.这一现象称为**杠杆效应**(leverage effect).但卖权不一定有杠杆效应.

2.2.3 未定权益定价和复制的鞅方法

本节在布莱克-索尔斯模型框架下介绍如何用鞅方法来研究未定权益的定价和复制.令$(\mathscr{F}_t)$为布朗运动(B_t)的自然 σ 代数流.设$\{a,b\}$为一交易策略,(V_t)为它的财富过程,$(\widetilde{V}_t)$为其贴现过程,为要该交易策略是自融资的,则必须且只需

$$d\widetilde{V}_t = a(t)d\widetilde{S}_t, \tag{2-21}$$

其中 $\widetilde{S}_t = e^{-rt}S_t$.(2-13)式可改写为

$$d\ \tilde{S}_t = \tilde{S}_t[(\mu - r)dt + \sigma dB_t].$$

于是,若由

$$\left.\frac{dP^*}{dP}\right|_{\mathscr{F}_T} = \exp\left[-\frac{\mu - r}{\sigma}B_T - \frac{1}{2}\left(\frac{\mu - r}{\sigma}\right)^2 T\right] \tag{2-22}$$

定义一概率测度 P^*,则由基尔沙诺夫(I. V. Girsanov)定理知,$B_t^* = B_t + \frac{\mu - r}{\sigma}$为一 P^*布朗运动,且有

$$d\ \tilde{S}_t = \tilde{S}_t\ \sigma dB_t^*.$$

于是$(\tilde{S}_t)$为一 P^*鞅．由(2-21)式知,任一自融资可容许交易策略的贴现财富过程为一非负局部鞅,从而为一上鞅．由此立刻推知市场无套利,因为零初值非负上鞅恒等于零．

称 P^*为市场的等价鞅测度．由(2-13)式得

$$dS_t = S_t[rdt + \sigma dB_t^*]. \tag{2-23}$$

这表明在 P^*下股票的预期收益率等于无风险证券利率 r．因此,P^*也称为风险中性概率测度．(2-13)式还可写成

$$dS_t = S_t[(r + \sigma\eta)dt + \sigma dB_t],$$

其中 $\eta = \frac{\mu - r}{\sigma}$,称 η 为股票的风险市场价格．

下一定理是鞅方法用于未定权益的定价和复制的主要结果．

定理 3　令 ξ 为一欧式未定权益,在 P^*下可积．则存在一复制 ξ 的容许的自融资交易策略$\{a, b\}$,使得其财富过程为

$$V_t = E^*[e^{-r(T-t)}\xi \mid \mathscr{F}_t], \tag{2-24}$$

即其贴现财富过程$(\tilde{V}_t)$为一 P^*鞅．此外,复制 ξ 的自融资交易策略$\{a, b\}$是唯一的．如果 $\xi = f(S_T)$,$V_t = F(t, S_t)$且 $F \in C^{1,2}([0, T) \times R_+)$,则 $a(t) = F_x(t, S_t)$．

注意,可把 V_t 定义为未定权益 ξ 在时刻 t 的公平价格(fair price),因为这是唯一的无套利价格．称由(2-24)式给出的定价为套利定价或风险中性定价．

定理 4　如果 $\xi = f(S_T)$,则 $V_t = F(t, S_t)$,其中

$$F(t, x) = e^{-r(T-t)}\int_{-\infty}^{\infty} f\left(xe^{(r-\sigma^2/2)(T-t)+\sigma y\sqrt{T-t}}\right)\frac{e^{-y^2/2}}{\sqrt{2\pi}}dy. \tag{2-25}$$

由(2-25)式可以直接推得布莱克-索尔斯公式,而不需要解布莱克-索尔斯方程．

2.2.4 美式期权的定价

假定市场模型是布莱克-索尔斯模型．美式未定权益是一$(\mathscr{F}_t)$适应过程(Z_t)．它的(卖方)价格过程为

$$V_t = \operatorname*{ess.sup}_{\tau \in \mathscr{T}_{t,T}} E^*[e^{-r(T-\tau)}Z_\tau \mid \mathscr{F}_t]. \tag{2-26}$$

其中 $\mathcal{T}_{t,T}$为取值于$[t,T]$的停时，ess. sup 为本性上确界．可以证明，美式买权(即 $Z_t=(S_t-K)^+$)的定价和欧式买权的定价是相同的．对美式卖权(即 $Z_t=(K-S_t)^+$)，有 $V_t=\Phi(t,S_t)$，其中

$$\Phi(t,x)=\sup_{\tau\in\mathcal{T}_{0,T-t}} E(e^{r\tau}K-xe^{\sigma B_\tau-\frac{1}{2}\sigma^2\tau})^+. \tag{2-27}$$

如果在停时 τ 执行美式卖权合约，则它在时刻 0 的价值为

$$V_0^\tau=E^*[e^{-r\tau}(K-S_\tau)^+].$$

令

$$s^*(t)=\sup\{x\geqslant 0;\Phi(t,x)=(K-x)^+\},\quad t\leqslant T,$$

则 s^* 为$[0,T]$上的 C^∞ 非降函数，且$\lim\limits_{t\to T}s^*(t)=K$. 可以证明，若令

$$\tau^*=\inf\{t\in[0,T];S_t=s^*(t)\},$$

当且仅当在时刻 $\tau=\tau^*$时执行合约才使 V_0^τ达最大．换言之，令 $\mathscr{D}=\{(t,x);t<T,x>s^*(t),\}$，当$(t,S_t)$首次从内部穿过 $\mathscr{D}$的边界时执行合约是最优的．因此，称$s^*(t)$为时刻 t 的股票**临界价格**(critical price)，称 $\mathscr{D}$ 为**不执行区域**(contnuation region). 可以证明，在区域 $\mathscr{D}$ 内，$\Phi(t,x)$满足布莱克-索尔斯方程，在边界$(t,s^*(t))$上满足如下条件：

$$\Phi(t,s^*(t))=(K-s^*(t))^+,\quad \frac{\partial\Phi}{\partial t}(t,s^*(t))=-1.$$

因此，美式卖权的定价归结为偏微分方程中的一个自由边界问题．$s^*(t)$预先是不知道的，称为**自由边界**(free boundary). 在实际应用中，能够给出自由边界的一个较好的估计很重要．一个理论结果是：当 $t\to T$，渐近地有 $K-s^*(t)\sim K\sigma\sqrt{(T-t)\log(1/(T-t))}$.

2.3 特异期权的定价

不是通常的买权或卖权的期权统称为**特异期权**(exotic option). 在金融市场中最常见的特异期权有**障碍期权**(barrier option)、**亚式期权**(Asian option)和**回看期权**(lookback option). 这些期权常被公司或金融机构用于风险管理．本节介绍它们的定价，且采用上一节中的布莱克-索尔斯模型及记号．

2.3.1 障碍期权

顾名思义，所谓障碍期权就是预先对期权的标的资产的价位设置(单边或双边的)界限．如果在期权到期之前标的资产的价格穿越界限，则期权价值变成零，否则到期按通常期权结算．例如，**双敲期权**(double-knock option)，它有两个障碍 L,U，对买权或卖权，需假定约定价 K 满足 $L<K<U$. 设未设障碍时的期权权益为 $g(S_T)$，则双敲期权的价格过程为 $F(t,S_t)$，其中 $F(t,x)$满足布莱克-索尔斯方程(2-15)，边界条件为

$$F(t,L)=F(t,U)=0,\quad t<T;\quad F(T,x)=g(x),\quad L<x<U.$$

其解可以表成一无穷级数．对单障碍期权，可以得到显式解．例如，考虑下失效买权(down-and-out call option)，其约定价为 K，到期时刻为 T，障碍 $X < K$．假定期权的标的资产为股票，其价格过程 (S_t) 服从布莱克-索尔斯模型．可以证明，期权在时刻 t 的价格为 $\widetilde{C}(t, S_t)$，其中

$$\widetilde{C}(t,x) = C(t,x) - \left(\frac{x}{X}\right)^{-(k_1-1)} C(t, X^2/x), \tag{2-28}$$

其中，$k_1 = \frac{2r}{\sigma^2}$，$C(t,x)$ 由布莱克-索尔斯公式(2-16)给出．

2.3.2 亚式期权

亚式期权的权益依赖于期权的标的资产价格在期权有效期限内的几何或算术平均值．如果用平均价格代替普通期权中的约定价，称这种期权为平均约定期权(average strike option)．如果用平均价格代替普通期权中的标的资产在期权到期时的价格，称这种期权为平均标价期权(average rate option)．

下面只考虑几何平均标价买权，其权益为

$$\xi = \left(\exp\left(\frac{1}{T}\int_0^T \log(S_u)\,\mathrm{d}u\right) - K\right)^+. \tag{2-29}$$

用 C_t 表示 ξ 在时刻 t 的价格．令 P^* 为等价鞅测度，则由风险中性定价原理，有

$$C_t = \mathrm{e}^{-r(T-t)} E^*[\xi \mid \mathscr{F}_t]. \tag{2-30}$$

通过计算得到 $C_t = \mathrm{e}^{-r(T-t)} F(t, X_t)$，其中

$$F(t,x) = \mathrm{e}^{x r^*(T-t)} \int_{-\infty}^{\infty} \left(\mathrm{e}^{\sigma^* \sqrt{T-t}\,y} - K\mathrm{e}^{-r^*(T-t)}\right)^+ \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\,\mathrm{d}y$$

$$= x\mathrm{e}^{\left(r^* + \frac{\sigma^{*2}}{2}\right)(T-t)} N(d_1^*) - KN(d_2^*), \tag{2-31}$$

$$\begin{cases} d_1^* = \dfrac{\log(x/K) + (r^* + \sigma^{*2})(T-t)}{\sigma^* \sqrt{T-t}}; \\ d_2^* = \dfrac{\log(x/K) + r^*(T-t)}{\sigma^* \sqrt{T-t}}. \end{cases} \tag{2-32}$$

对算术平均标价期权的定价，得不到显式解．

2.3.3 回看期权

回看期权的权益依赖于期权有效期限内的标的资产价格的最大或最小值．如果用最小值(最大值)代替买权(相应地，卖权)中的约定价，称这种期权为回看约定期权(lookback strike option)．如果用最大值(最小值)代替买权(相应地，卖权)中的标的资产到期价格，称这种期权为**回看标价期权**(lookback rate option)．

回看约定买权(相应地，卖权)的权益为

$$\xi = S_T - \min_{0 \leqslant s \leqslant T} S_s, \quad (\text{相应地}, \eta = \max_{0 \leqslant s \leqslant T} S_s - S_T).$$

用 C_t 和 P_t 分别表示它们在时刻 t 的价格，则有

$$\begin{cases} C_t = e^{-r(T-t)} E^* [\xi | \mathscr{F}_t]; \\ P_t = e^{-r(T-t)} E^* [\eta | \mathscr{F}_t]. \end{cases} \tag{2-33}$$

利用概率论中的公式

$$P(\max_{s \leq t}(\sigma B_s + \lambda s) \leq x) = N\left(\frac{x - \lambda t}{\sigma\sqrt{t}}\right) - e^{2\lambda x/\sigma^2} N\left(\frac{-x-\lambda t}{\sigma\sqrt{t}}\right), \tag{2-34}$$

其中 $x \geq 0$，$N(z)$为标准正态分布．可推得

$$P_t = S_t(-1 + N(d_3)(1 + \sigma^2/2r)) + M_t e^{-r(T-t)}\left(N(d_1) - \frac{\sigma^2}{2r}(S_t^{-1}M_t)^{(2r/\sigma^2)-1}N(d_2)\right), \tag{2-35}$$

其中 $M_t = \max_{0 \leq s \leq t} S_s$，

$$d_1 = \frac{\log(M_t/S_t) - (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}};$$

$$d_2 = \frac{-\log(M_t/S_t) - (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}};$$

$$d_3 = \frac{-\log(M_t/S_t) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}.$$

类似地，由

$$P(\min_{s \leq t}(\sigma B_s + \lambda s) \leq -x) = N\left(\frac{-x-\lambda t}{\sigma\sqrt{t}}\right) + e^{-2\lambda x/\sigma^2} N\left(\frac{-x+\lambda t}{\sigma\sqrt{t}}\right),$$

可得 C_t 的明显表达式．

回看标价买权的权益为

$$\xi = (\max_{0 \leq s \leq T} S_s - K)^+,$$

它在时刻 t 的价格为

$$C_t = S_t N(d_3)(1 + \sigma^2/2r) + K_t e^{-r(T-t)}\left(N(d_1) - \frac{\sigma^2}{2r}(S_t^{-1}K_t)^{(2r/\sigma^2)-1}N(d_2)\right) - e^{-r(T-t)}K, \tag{2-36}$$

其中 $K_t = \max(M_t, K)$，

$$d_1 = \frac{\log(K_t/S_t) - (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}};$$

$$d_2 = \frac{-\log(K_t/S_t) - (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}};$$

$$d_3 = \frac{-\log(K_t/S_t) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}.$$

2.4 利率的期限结构模型

在布莱克-索尔斯模型中，曾假定无风险证券的利率为一常数，这对考虑股票一类的短期期权的定价问题尚可接受，但若要考虑利率衍生产品的定价，这便是一个不可接受的假定．于是必须考虑随机利率．基本上有两类不同的方法研究利率的期限结构：一是短期利率模型，二是远期利率模型．本节首先介绍几种常见的短期利率模型和一种远期利率模型，然后介绍利率衍生产品的定价．

2.4.1 债券市场

以下固定一时间区段$[0,T]$．令(B_t)为一概率空间$(\Omega,\mathscr{F},P)$上的d维布朗运动，其自然σ代数流记为$(\mathscr{F}_t)$．

考虑一债券市场，它由银行账户和各种不同期限的**贴现债券**(discount bond)或**零息债券**(zero-coupon bond)组成．后者指一种有价证券，它的售价低于它到期的票面价值，中间不付息．下面简称贴现债券为债券，并称到期时刻为s的债券为s债券．s债券在时刻$t\leqslant s$的价格记为$P(t,s)$．今后恒假定债券的票面价均为1(银行账户的计价单位)，于是有$P(s,s)=1$．

s债券在时刻t的到期收益率(yield-to-maturity)，简称收益率(yield)，定义为

$$Y(t,s)=-\frac{\log(P(t,s))}{s-t}. \tag{2-37}$$

它是在时刻t对利率的未来值的一个度量．在时刻t的**收益率曲线**(yield curve)是债券收益率$Y(t,s)$随到期时刻s变化的图像，它作为s-t的函数称为**利率的期限结构**(term structure of interest rates)．在时刻t的**短期利率**(short rate)$r(t)$定义为

$$r(t)=\lim_{s>t,s\to t}Y(t,s),$$

若该极限存在．下面假定对每个$t\in[0,T]$，$r(t)$存在并且过程$(r(t))$有可测修正．此外假定$\int_0^T r(t)\mathrm{d}t<\infty$．

如果$P(t,s)$关于s可微，则可用所谓的**远期利率**(forward rates)$f(t,s)$来对利率的未来值进行度量，其定义为

$$f(t,s)=-\frac{\partial\log P(t,s)}{\partial s}=-\frac{(\partial P(t,s))/\partial s}{P(t,s)}. \tag{2-38}$$

一旦知道了远期利率$f(t,s)$，可以重新由下式求得债券价格$P(t,s)$：

$$P(t,s)=\exp\left(-\int_t^s f(t,u)\mathrm{d}u\right). \tag{2-39}$$

权益依赖于未来的利率或债券价格的金融合约称为**利率衍生产品**(interest rate derivative)．为了给利率衍生产品定价，需要对利率或债券价格的变化规律建立模型．基本假定是在不同到期的债券之间不存在套利机会．如果对一切$0\leqslant s\leqslant T$，$P(t,s)_{t\leqslant s}$为确定性函数，则在无套利假定下，必有

$$P(t,s)=\exp\left(-\int_t^s r(u)\mathrm{d}u\right).$$

特别地,这时债券价格由短期利率完全确定. 但在随机环境下,这一事实不再成立.债券价格一般由短期利率和一等价鞅测度确定. 事实上,设 $r(t)$为一非负可测适应过程,P^*为一等价鞅测度,则由风险中性定价原理,债券价格为

$$P(t,s)=E^*[\exp(-\int_t^s r(u)\mathrm{d}u)|\mathscr{F}_t],\quad t\leqslant s\leqslant T. \tag{2-40}$$

2.4.2 短期利率模型

首先考虑单因子模型,即假定在概率测度 P 下,$(r(t))$为一扩散过程:

$$\mathrm{d}r(t)=\mu_0(t,r(t))\mathrm{d}t+\sigma(t,r(t))\mathrm{d}B_t,\ t\leqslant T, \tag{2-41}$$

其中,(B_t)为一维标准布朗运动. 设 P^*为一与 P 等价的概率测度,则由(2-38)式定义的 $P(t,s)$可以作为一无套利市场中 s 债券在时刻 t 的价格. 选择不同的等价概率测度能导致不同的债券价格. 下面将看到,等价概率测度的选择归结为规定市场的风险价格. 为简单起见,下面只考虑那些关于 P 的拉东-尼古丁导数具有如下形式的 P^*:

$$\frac{\mathrm{d}P^*}{\mathrm{d}P}=\exp\left\{-\int_0^T\lambda(u,r(u))\mathrm{d}B_u-\frac{1}{2}\int_0^T\lambda^2(u,r(u))\mathrm{d}u\right\}, \tag{2-42}$$

其中 $\lambda(t,x)$为$[0,T]\times\mathbf{R}$上的一波雷尔(Borel)可测函数. 这时选择等价概率测度 P^*在于确定函数 λ,而 $\lambda(t,r(t))_{0\leqslant t\leqslant T}$正是风险的市场价格. 在风险中性概率 P^*下,

$$\mathrm{d}r(t)=\mu(t,r(t))\mathrm{d}t+\sigma(t,r(t))\mathrm{d}B_t^*,\quad t\leqslant T, \tag{2-43}$$

其中 $\mu(t,x)=\mu_0(t,x)-\sigma(t,x)\lambda(t,x)$, $B_t^*=B_t+\int_0^t\lambda(u,r(u))\mathrm{d}u$ 在 P^* 下为一标准布朗运动. 这时有 $P(t,s)=F(t,r(t);s)$,其中对一切 $s\in(0,T]$, $F(\cdot,\cdot;s)$为$[0,s]\times\mathbf{R}$上的一 $C^{1,2}$函数,且 $F(t,x;s)$为如下方程的唯一解:

$$F_t(t,x;s)+\mu(t,x)F_x(t,x;s)+\frac{1}{2}\sigma(t,x)F_{xx}(t,x;s)-xF(t,x;s)=0, \tag{2-44}$$

终端条件为 $F(s,x;s)=1$.

瓦西赛克(O.A.Vasicek)模型是第一个单因子模型(1977 年提出). 该模型假定短期利率 $r(t)$ 在风险中性概率 P^* 下为一奥因斯坦-乌伦贝克(Ornstein-Uhienbeck)过程,即满足

$$\mathrm{d}r(t)=a(b-r(t))\mathrm{d}t+\sigma\mathrm{d}B_t^*, \tag{2-45}$$

其中 a,b,σ 为正常数,(B_t^*)在 P^*下为一标准布朗运动.(2-45)式的解为

$$r(t)=r(0)\exp(-at)+b[1-\exp(-at)]+\sigma\exp(-at)\int_0^t\exp(as)\mathrm{d}B_s^*. \tag{2-46}$$

这时,s 债券的价格为

$$P(t,s)=\exp(A(t,s)-B(t,s)r(t)), \tag{2-47}$$

其中

$$B(t,s)=\frac{1-\exp(-a(s-t))}{a},\tag{2-48}$$

$$A(t,s)=\frac{(B(t,s)-st)(a^2b-\sigma^2/2)}{a^2}-\frac{\sigma^2B(t,s)^2}{4a}.\tag{2-49}$$

由(2-46)式知,$r(t)$服从正态分布,从而以正概率取负值,这显然不合理.为了克服瓦西赛克模型的这一缺点,考克斯、英格绍尔(J. E. Ingersoll)和罗斯于1985年建议用如下的短期利率模型(称为 **CIR 模型**):

$$dr(t)=a(b-r(t))dt+\sigma\sqrt{r(t)}dB_t^*.\tag{2-50}$$

这时 s 债券的价格仍由(2-47)式给出,不同的是,式中

$$B(t,s)=\frac{2[\exp(\gamma(s-t))-1]}{(\gamma+a)[\exp(\gamma(s-t))-1]+2\gamma},\tag{2-51}$$

$$A(t,s)=\frac{2ab}{\sigma^2}\lg\left[\frac{2\gamma[\exp(a+\gamma)(s-t)/2]}{(\gamma+a)[\exp(\gamma(s-t))-1]+2\gamma}\right],\tag{2-52}$$

其中 $\gamma=\sqrt{a^2+2\sigma^2}$.

在上述两个模型中,$B(t,s)$和$A(t,s)$都是 s 和 t 的确定性函数,收益率曲线 $Y(t,s)$为短期利率 $r(t)$的线性函数:

$$Y(t,s)=\frac{1}{s-t}[B(t,s)r(t)-A(t,s)],$$

因此称这两个模型具有仿射期限结构(affine term structure).

在实际应用中,需用短期利率的历史数据来估计参数 b,a 和σ,然后基于这些参数计算债券的价格,并与当前市场价格比较,再对参数进行调整,以使模型尽量拟合债券价格的历史观测值.但这两个模型都难以做到很好拟合当前的债券价格.为克服这一缺点,霍尔(J. E. Hull)和怀特(P. A. White)于1990年将上述两个模型推广到时变系数的情况.

$$dr(t)=(\Phi(t)-a(t)r(t))dt+\sigma(t)dB_t^*,$$

$$dr(t)=(\Phi(t)-a(t)r(t))dt+\sigma(t)\sqrt{r(t)}dB_t^*.$$

这时模型仍具有仿射期限结构.

上面介绍的单因子模型并不能很好拟合真实利率的变动情况.现在比较流行的且在数学上容易处理的多因子模型,是一种所谓的高维平方高斯马尔科夫过程,它由如下方程描述:

$$dX_t=(a(t)+C_tX_t)dt+\sigma_t dB_t^*,$$

$$r(t)=\frac{1}{2}|X_t|^2,$$

其中(B_t^*)为一 P^* 下的 d 维布朗运动,σ,C 为 $\mathrm{R}^d\times\mathrm{R}^d$ 值函数,a 为 R^d 值函数.这一模型的好处是,它也能给出债券价格的明显表达式.

2.4.3 HJM 模型

1987年赫斯(D. Heath)、贾鲁(A. Jarrow)和毛顿(A. Morton)建议直接用远期利率模型来描述利率的期限结构.他们假定在风险中性概率下,对每个固定的 $s\leqslant T$,

远期利率作为 t 的函数为一伊藤过程：

$$f(t,s)=f(0,s)+\int_0^t\mu(u,s)\mathrm{d}u+\int_0^t\sigma(u,s)\mathrm{d}B_u^*,\quad t\leqslant s,\tag{2-53}$$

其中 (B_t^*) 为一 P^* 下的 d 维布朗运动，$\mu(\cdot,s),\sigma(\cdot,s)$ 分别为 $\mathbf{R}$ 值和 $\mathbf{R}^d$ 值适应可测过程．初始远期利率 $f(0,s)$ 是一确定性函数，满足 $\int_0^T f(0,u)\mathrm{d}u<\infty$．由市场无套利假定，在一定的技术性条件下，可以证明：$\mu(t,s)$ 必须满足

$$\mu(t,s)=\sigma(t,s)\cdot\int_t^s\sigma(t,u)\mathrm{d}u\ .\tag{2-54}$$

这时，有

$$r(t)=f(0,t)+\int_0^t\sigma(v,t)\cdot\int_v^t\sigma(v,u)\mathrm{d}u\mathrm{d}v+\int_0^t\sigma(v,t)\mathrm{d}B_v^*.$$

特别地，若 $\sigma(t,s)$ 为一常数，则得到何(T.S.Ho)-李(S.B.Lee)模型的连续极限．

$$\mathrm{d}r(t)=\Phi(t)\mathrm{d}t+\sigma\mathrm{d}B_t^*,$$

其中 $\Phi(t)=\sigma^2t+\dfrac{\partial f(0,t)}{\partial t}$.

2.4.4 利率衍生产品的定价

在研究利率衍生产品的定价时，对计价单位(numeraire)有两种可能选择，一是银行账户，二是 T-债券．如果衍生产品的标的状态变量为短期利率且利率模型为单因子的伊藤随机微分方程(2-41)(例如，瓦西赛克模型或 CIR 模型)，则可选银行账户作为计价单位．假定某衍生产品的到期时刻 $\tau\leqslant T$，在 τ 以前连续派息率为 $h(t,r(t))$，到期支付 $g(\tau,r(\tau))$，则由风险中性定价原理，衍生产品在时刻 t 的价格为

$$F(t,r(t))=E^*[\int_t^\tau\phi_{t,s}h(s,r(s))\mathrm{d}s+\phi_{t,\tau}g(\tau,r(\tau))\mid\mathscr{F}_t],\tag{2-55}$$

其中 $\phi_{t,s}=\exp[-\int_t^s r(u)\mathrm{d}u]$．在一定条件下，由概率论中的费因曼-卡茨公式(Feynman-Kac formula)知，F 为如下偏微分方程的解：

$$\mathscr{D}F(t,x)-xF(t,x)+h(t,x)=0,\quad(t,x)\in[0,\tau)\times\mathbf{R}^d,\tag{2-56}$$

边界条件为

$$F(\tau,x)=g(\tau,x),\quad x\in\mathbf{R}^d.\tag{2-57}$$

这里

$$\mathscr{D}F(t,x)=F_t(t,x)+F_x(t,x)\mu(t,x)+\frac{1}{2}F_{xx}(t,x)\sigma(t,x)^2.\tag{2-58}$$

特别地，τ 债券在时刻 t 的价格为 $P(t,\tau)=f(t,r(t))$，其中 f 为方程(2-56)($h=0$)的解，边界条件为 $f(\tau,x)=1$.

现在假定远期利率服从 **HJM 模型**．这时可取 T 债券作为计价单位．更确切地，令 $\alpha_t=P(t,T)/P(0,T)$，取 (α_t) 作为计价单位．设 P^* 为选银行账户作为计价单位时的等价鞅测度，即债券的贴现价格过程在 P^* 下为一鞅，若令 Q 为 $(\Omega,\mathscr{F}_T)$

上的概率测度，使得

$$\frac{dQ}{dP^*}=\frac{\alpha_T}{\beta_T}=\frac{1}{P(0,T)\beta_T},$$

则有

$$L_t=E^*\left[\frac{dQ}{dP^*}\mid\mathscr{F}_t\right]=\frac{P(t,T)}{P(0,T)\beta_t}=\frac{\alpha_t}{\beta_t},$$

从而(β_t/α_t)为一 Q 鞅．这表明 Q 为取(α_t)作为计价单位时的等价鞅测度．设 ξ 是到期时刻为 T 的未定权益，V_t 为它在时刻 t 时的价格，则有

$$V_t=\beta_tE^*[\beta_T^{-1}\xi\mid\mathscr{F}_t]=P(t,T)E_Q[\xi\mid\mathscr{F}_t].\qquad(2\text{-}59)$$

3　动态投资组合理论

第1章考虑的单期间最优投资组合问题，可以看成是投资者在期间末消费他的所有财富，而他的投资目标是使财富的期望效用达到最大．这种简化显然对一般的投资决策是不令人满意的．比较接近现实的是考虑多期间的投资决策，即研究所谓的跨期(intertemporal)的投资组合问题，这样可以动态地考虑消费和投资．在研究动态跨期模型时，通常采用连续时间模型，因为它虽然用到较深的数学工具，但较离散时间模型更能反映动态特性，且能得到更精确的理论结果．

首先不考虑消费，只讨论一个简单的动态投资组合问题：投资者希望找一个自融资交易策略，使他在某个预定的时刻 T 的财富的期望效用达到最大．假定市场模型是布莱克-索尔斯模型，即假定市场上只有两种证券，一是风险证券(如股票)，二是银行账户，连续复利率为 r，股票价格过程(S_t)为几何布朗运动(见(2-13)式)．假定$(\mathscr{F}_t)$为布朗运动的自然 σ 代数流．令 $H^2(P)$表示那些满足 $E[\int_0^T|a(t)\times S_t|^2dt]<\infty$ 的自融资交易策略$\{a,b\}$的全体，$L_+^2(P)$为 $\mathscr{F}_T$-可测的非负随机变量全体．由伊藤随机分析中的一个基本结果(鞅表示定理)推知：任何属于 $L_+^2(P)$的未定权益都可以用 $H^2(P)$中的一自融资交易策略来复制．

设投资者的初始财富为 W_0，他的投资目标是使他在时刻 T 的财富的期望效用最大化．设 V 是他的效用函数，假定 V 为严格凹的非降可微函数．投资者寻找最优动态投资组合$\{a^*,b^*\}$归结为解如下动态规划问题：

$$\begin{cases}\sup\limits_{\{a,b\}\in H^2(P)} E[V(W_T)],\\ a(0)S_0+b(0)=W_0,\end{cases}\qquad(3\text{-}1)$$

令 P^* 为市场的等价鞅测度(即 P^* 由(2-22)式确定)，则由未定权益定价的鞅方法容易推知，$\{a^*,b^*\}$为(3-1)式的解，等价于 $W_T^*=a^*(T)S_T+b^*(T)e^{rT}$为如下静态极大值问题的解：

$$\begin{cases}\sup\limits_{X\in L_+^2(P)} E[V(X)];\\ E^*[e^{-rT}X]=W_0.\end{cases}\qquad(3\text{-}2)$$

令

$$\eta(t)=\exp\left(-\frac{\mu-r}{\sigma}B_t-\frac{1}{2}\left(\frac{\mu-r}{\sigma}\right)^2\right),\quad 0\leqslant t\leqslant T.$$

由于$\frac{\mathrm{d}P^*}{\mathrm{d}P}\Big|_{\mathscr{F}_T}=\eta(T)$,故有

$$E^*[\mathrm{e}^{rT}W_T^*]=E[\mathrm{e}^{-rt}\eta(T)W_T^*].\tag{3-3}$$

利用拉格朗日乘子法可以证明:设 f 为 V' 的反函数,即

$$f(x)=\inf[z\geqslant 0:V'(z)\leqslant x],$$

且令 $\xi(t)=\lambda\mathrm{e}^{-rt}\eta(t)$,则(3-2)式的解可表为 $W_T^*=f(\xi(T))$,其中 $\lambda>0$ 为一适当常数(拉格朗日乘子). 进一步由此可以证明(3-1)式的解为$\{a^*,b^*\}$,其中

$$\begin{cases}a^*(t)=-\dfrac{\mu-r}{\sigma^2S(t)}F_x(t,\xi(t))\xi(t);\\ b^*(t)=[F(t,\xi(t))-a^*(t)S(t)]\mathrm{e}^{-rt}.\end{cases}\tag{3-4}$$

$F(t,x)$为下述偏微分方程的解

$$\begin{cases}\dfrac{(\mu-r)^2}{2\sigma^2}F_{xx}x^2-F_x\left(r-\dfrac{(\mu-r)^2}{\sigma^2}\right)x-rF+F_t=0;\\ F(0,\lambda)=W_0.\end{cases}\tag{3-5}$$

注意,与期权定价的布莱克-索尔斯方程(2-15)显著不同的是,在方程(3-4)中出现了股票的预期收益率 μ. 因此,如果投资者不知道 μ,他就无法找到他的最优动态投资组合策略(即交易策略).

下面考虑投资者的最优消费和投资组合问题 . 设投资者的初始财富为 W_0,在时刻 t 的消费效用函数为$u(t,x)$,在终了时刻 T 时剩余财富的效用函数为 $V(x)$. 投资者的目标是使消费和终了财富的期望效用之和达最大 . 用 $C(P)$表示满足 $E\left[\int_0^T c^2(t)\mathrm{d}t\right]<\infty$的非负适应过程 $c(t)$的全体,则问题成为寻找带消费的可容许的投资组合$\{a^*,b^*,c\}$(见下面的(3-7))式,使之成为如下的动态规划问题的解:

$$\sup_{(a,b,c)\in H^2(P)\times C(P)}E\left[\int_0^T u(t,c(t))\mathrm{d}t+V(W_T)\right],\tag{3-6}$$

其中

$$W_t=a(t)S_t+b(t)\mathrm{e}^{rt}=W_0+\int_0^t a(s)\mathrm{d}S_s+r\int_0^t b(s)\mathrm{e}^{rs}\mathrm{d}s-\int_0^t c(s)\mathrm{d}s,\tag{3-7}$$

$c(t)$表示在时刻 t 的消费率 . 该问题等价于如下的静态极大值问题:

$$\begin{cases}\sup\limits_{(c,X)\in C\times L_t^2(P)}E\left[\int_0^T u(t,c(t))\mathrm{d}t+V(X)\right];\\ E^*\left[\int_0^T \mathrm{e}^{-rt}c(t)\mathrm{d}t+\mathrm{e}^{-rT}X\right]=W_0.\end{cases}\tag{3-8}$$

与(3-2)式类似,利用拉格朗日乘子法可以解此问题 . 事实上,设 f 和 $g(t,\cdot)$分别为 V' 和 $u_x(t,\cdot)$的反函数,即

$$f(x)=\inf\{z\geqslant 0:V'(z)\leqslant x\},$$

$$g(t,y)=\inf\{x\geq 0: u_x(t,x)\leq y\},$$

令 $\xi(t)=\lambda e^{-rt}\eta(t)$，则(3-8)式的解可表为

$$c^*(t)=g(t,\xi(t)),\quad W_T^*=f(\xi(T)). \tag{3-9}$$

其中 $\lambda>0$ 为一适当常数(拉格朗日乘子)．进一步由此可以证明(3-6)式的解为 $\{a^*,b^*,c^*\}$，其中

$$a^*(t)=F_y(t,\xi(t),S_t)+\frac{\mu-r}{\sigma^2 S(t)}F_x(t,\xi(t),S_t)\xi(t), \tag{3-10}$$

$$b^*(t)=[F(t,\xi(t),S_t)-a^*(t)S(t)]e^{-rt},$$

$c^*(t)$由(3-9)式给出，$F(t,x,y)$为下述偏微分方程的解：

$$\begin{cases}\frac{(\mu-r)^2}{\sigma^2}F_{xx}x^2+\frac{\sigma^2}{2}F_{yy}y^2-\frac{\mu-r}{2}F_{xy}xy-\\(r-\frac{(\mu-r)^2}{\sigma^2})F_x x+F_y ry-rF+F_t+g(t,x)=0;\\F(T,x,y)=f(x),F(0,S_0,\lambda)=W_0.\end{cases} \tag{3-11}$$

参考文献

1 Black F ,Scholes M.The pricing of options and corporate liabilities. J of Political Economy,1973(81):635 ~ 654

2 Cox J, Ross S, Rubinstein M. Option pricing: a simplified approach J Fin Econ.,1979 (7):229 ~ 263

3 Duffie D. Dynamic asset pricing theory. 2nd ed. New Jersey: Princeton Univ Press, 1996.

4 Huang C F, Litzenberger R H. Foundations for financial economics. Amsterdam: North-Holland, 1988.

5 Karatzas I. Lectures on the mathematics of finance. CRM Monograph Series, Vol 8 American Mathematical Society, Providence, Phode Island, USA, 1997.

6 Merton R C. Theory of rational option pricing. Bell J Econ. and Manag Sci, 1973(4): 141 ~ 183

7 Merton R C. Continuous-Time finance, basil blackwell. Cambridge: Oxford, 1990.

8 Musiela M Rutkowski M. Martingale methods in financial modelling. Milan: Springer. 1997.

9 Wilmott P, Dewynne, J., Howison S. Option pricing: mathematical models and computations. Cambridge: Oxford Financial Press, 1993.

10 Yan J A. Introduction to martingale methods in option pricing. LN in Math 4, Liu Bie Ju Center for Math Sciences, City Univ of Hong Kong, 1998.

11 叶中行，林建忠．数理金融——资产定价与金融决策理论．北京：科学出版社，1998．

·经济数学卷·

第 4 篇

经济控制论

编　者　张金水
审校者　陈叔平

目　录

引　　言

经济控制论是将系统论思想、控制理论知识与经济学紧密结合在一起的一门交叉性学科.经济控制论的研究可以从两方面入手:一是从控制理论的各个基本分支出发,将经济系统作为背景来研究,相应地,经济控制论可划分为确定性动态系统经济控制论、随机动态系统经济控制论、广义系统经济控制论、非线性动态系统经济控制论,等等;二是以经济理论为主线,综合应用系统论思想、控制理论知识来描述经济学,并研究其发展与应用.

自从维纳1948年创建了控制论这门新学科之后,将控制论应用于经济管理领域的论文与专著随之大量面世.其中波兰经济学家兰格(O.Lange)、罗马利亚经济学家曼内斯库(M.Manescu)、美籍华人经济学家邹至庄等人的专著具有广泛的影响.在我国,张仲俊及乌家培等人在经济控制论领域做出了开创性工作.经济控制论目前正处在不断发展之中.

经济控制论是数理经济学的一个特殊分支.目前习惯上把经济系统运动稳定性分析、经济系统运动的目标设定、市场运动的能控性与调控政策设计、经济系统在一定时间内到达约定状态的能达性等归结为经济控制论的研究内容.本篇将介绍其中最基本的内容.

经济控制论在经济管理各领域具有广泛的应用.例如在宏观经济管理中,可构造从经济策略变量到目标变量之间的因果关系模型,然后求解从当前状态到目标状态的最优发展轨道,求出运行在最优轨道上的产品价格、产出结构、经济增长率、利率、汇率、进出口量等变量的变化,以及应采取的相应策略.本篇介绍的内容是其中最基本的部分.

1　经济系统的运动分析

1.1　消费品需求函数

1.1.1　效用函数

设有 n 种消费品,当某人拥有各种消费品的量为 $x=[x_1,\cdots,x_n]^T$ 时,所得到的效用为 $U(x)$. $U(x)$ 称为效用函数.依经济学知识, $U(x)$ 应是凹函数.常用的效用函数数学表达式有

1.对数线性型

$$U(x)=A(x_1-a_1)^{\beta_1}(x_2-a_2)^{\beta_2}\cdots(x_n-a_n)^{\beta_n},\tag{1-1}$$

其中，$a_i \geqslant 0$ 为第 i 种消费品的最低需要量；A，β_i 为正常数，且 $\beta_1 + \cdots + \beta_n < 1$.

2. CES (constant elasticity of substitution) 型

$$U(x) = \left[\sum_{i=1}^{n} a_i^{1/\sigma} x_i^{(\sigma-1)/\sigma} \right]^{\delta/(\sigma-1)}, \tag{1-2}$$

其中，a_i，σ，$\delta(\delta < \sigma)$ 为正常数. 与上式类似的效用函数有

$$U(x) = \left[\sum_{i=1}^{n} a_i^{1/\sigma} (x_i - b_i)^{(\sigma-1)/\sigma} \right]^{\delta/(\sigma-1)}. \tag{1-3}$$

1.1.2 效用最大法则

若消费者消费支出总额为 M，那么在市场价格为 $p_1, \cdots, p_n$ 下购买各种消费品的数量由如下数学模型求解：

$$\begin{cases} \max & U(x), \\ \text{s.t.} & p_1 x_1 + \cdots + p_n x_n = M. \end{cases} \tag{1-4}$$

上式取极值的必要条件又称为**效用最大法则**，即

$$\begin{cases} \dfrac{\partial U / \partial x_i}{p_i} = \lambda, i = 1, \cdots, n; \\ p_1 x_1 + \cdots + p_n x_n = M. \end{cases} \tag{1-5}$$

其中，λ 为拉格朗日算子，经济含义为货币的边际效用.

1.1.3 需求函数

若给出效用函数 $U(x)$ 的具体数学表达式，通过求解(1-5)式，可求出消费品**需求函数**：

$$x_i = x_i(p_1, \cdots, p_n, M), \quad i = 1, \cdots, n. \tag{1-6}$$

例如，如果效用函数如(1-1)式所示，那么需求函数为

$$p_i x_i = p_i a_i + \frac{\beta_i}{\beta_1 + \cdots + \beta_n}(M - p_1 a_1 - \cdots - p_n a_n), \quad i = 1, \cdots, n. \tag{1-7}$$

上式所示的需求函数称为**线性支出系统**，即第 i 种消费品支出额 $p_i x_i$ 是价格及收入 M 的线性函数.

又如，如果效用函数如(1-2)式所示，那么需求函数为

$$x_i = \frac{a_i p_i^{-\sigma} M}{\sum_{i=1}^{n} a_i p_i^{1-\sigma}}, \quad i = 1, \cdots, n. \tag{1-8}$$

1.2 产品供给函数与要素需求函数

1.2.1 生产函数

生产函数反映产品生产过程中投入的生产要素数量与产品产出量之间的数量

关系.对不同类型的生产过程将对应不同的生产函数表达式.

常用的生产函数表达式有

1.柯布-道格拉斯(Cobb-Douglas)类型的生产函数

$$Y = A Z_1^{\beta_1} \cdots Z_m^{\beta_m}, \tag{1-9}$$

其中,Y 为产出量,Z_i 为投入的第 i 种要素量.当 $\beta_1 + \cdots + \beta_m = 1$ 时,称之为规模报酬不变的生产函数.如果投入的要素只有资本 K 与劳动工时 L 两种,那么(1-9)式通常写为

$$Y = AK^a L^b. \tag{1-10}$$

2. CES 类型

$$Y = \left(\sum_{i=1}^{m} a_i Z_i^{\sigma} \right)^{\delta/\sigma}, \tag{1-11}$$

其中,a_i, σ, δ 为常数,$a_i > 0$, $-\infty < \sigma < 1, \delta > 0$.当 $\delta = 1$ 时,(1-11)式为规模报酬不变的生产函数.

以上都属于投入要素可以互相替代的生产函数类型.

3.无联合产出且投入要素不可互相替代的生产函数

$$Y = \min\left\{ \frac{Z_1}{a_1}, \frac{Z_2}{a_2}, \cdots, \frac{Z_m}{a_m} \right\}, \tag{1-12}$$

其中,a_i 为投入系数,Z_i 为第 i 种要素实际投入量,Y 为产品产出量.无联合产出是指一种生产过程只生产一种产品.

4.有联合产出且投入要素不可互相替代的生产函数

$$Y = (y_1, y_2, \cdots, y_n) \times \min\left\{ \frac{Z_1}{a_1}, \frac{Z_2}{a_2}, \cdots, \frac{Z_m}{a_m} \right\}. \tag{1-13}$$

上式表明,当投入的要素恰好为 $Z_i = a_i, i = 1, \cdots, m$ 时,产出的 n 种产品量为 $y_i \geqslant 0, i = 1, \cdots, n$.

1.2.2 利润最大法则

对(1-9)~(1-11)式所示的投入要素可互相替代的生产函数来讲,如果各种要素价格为 $w_1, \cdots, w_m$,产品价格为 p,在不考虑加工时间的情况下,利润 Π 为

$$\Pi = pY - \sum_{i=1}^{m} w_i Z_i, \tag{1-14}$$

生产者利润最大的必要条件为

$$p \frac{\partial Y}{\partial Z_i} = w_i, \quad i = 1, \cdots, m. \tag{1-15}$$

上式又称为**利润最大法则**.

1.2.3 产品供给函数及要素需求函数

给出生产函数的数学表达式,依(1-15)式所示的利润最大法则,可求出产品供给函数及要素需求函数分别为

$$\begin{cases} Y = Y(p, w_1, \cdots, w_m), \\ Z_i = Z_i(p, w_1, \cdots, w_m), \quad i = 1, \cdots, m. \end{cases} \tag{1-16}$$

例如，如果生产函数如(1-9)式所示，那么依(1-15)式所示的利润最大法则，可得

$$p\beta_i Y / Z_i = w_i, \quad i = 1, \cdots, m.$$

从上式可求出单位产出时的要素需求函数

$$z_i = z_i(p, w_1, \cdots, w_m) \text{ 为}$$

$$z_i = z_i(p, w_1, \cdots, w_m) = Z_i / Y = p\beta_i / w_i. \tag{1-17}$$

将上式代入(1-9)式所示的生产函数，得

$$Y = A(p\beta_1 / w_1)^{\beta_1} \cdots (p\beta_m / w_m)^{\beta_m} Y^{\beta_1 + \cdots + \beta_m},$$

或

$$Y^{1-\beta_1-\cdots-\beta_m} = A \prod_{i=1}^{m} (\beta_i / w_i)^{\beta_i} \cdot p^{\beta_1 + \cdots + \beta_m}. \tag{1-18}$$

当生产函数为规模报酬递减，即 $\beta_1 + \cdots + \beta_m < 1$ 时，从(1-18)式可求出产品供给函数为

$$Y = \left\{ A \prod_{i=1}^{m} (\beta_i / w_i)^{\beta_i} \cdot p^{\beta_1 + \cdots + \beta_m} \right\}^{1/(1-\beta_1-\cdots-\beta_m)}. \tag{1-19}$$

要素需求函数可从(1-17)式及(1-19)式求出：

$$Z_i = (p\beta_i / w_i) \left\{ A \prod_{i=1}^{m} (\beta_i / w_i)^{\beta_i} \cdot p^{\beta_1 + \cdots + \beta_m} \right\}^{1/(1-\beta_1-\cdots-\beta_m)}. \tag{1-20}$$

当生产函数为规模报酬不变，即 $\beta_1 + \cdots + \beta_m = 1$ 时，从(1-18)式可求出按成本定产品价格的方程

$$p = p^* = \frac{1}{A} \prod_{i=1}^{m} (w_i / \beta_i)^{\beta_i}. \tag{1-21}$$

在这种情况下，(1-16)式所示的产品供给函数是不连续的函数，它可表述为

当产品市场价 $p > p^*$ 时，供给量 Y 尽可能大；

当产品市场价 $p = p^*$ 时，供给量 Y 等于市场需求量；

当产品市场价 $p < p^*$ 时，供给量 Y 为零.

1.3　供求平衡

设有 m 种要素及 n 种消费品，价格分别为 $w_1, \cdots, w_m$ 及 $p_1, \cdots, p_n$. 在其它条件不变的情况下，任一种要素或消费品的供给量与需求量可看做是价格的函数. 供给量 S_i 与需求量 D_i 的平衡由如下方程组描述：

$$S_i(p_1, \cdots, p_n, w_1, \cdots, w_m) = D_i(p_1, \cdots, p_n, w_1, \cdots, w_m), \tag{1-22}$$

其中，$i = 1, \cdots, n + m$.

讨论(1-22)式所示静态经济系统解的存在性与唯一性，按习惯归入数理经济学的可计算一般均衡分析范畴. 由于生产要素可以是资源、资本、劳动工时等，因此

(1-22)式包括了产品市场、资源市场、资本市场、劳动市场等各种市场的一般均衡分析.(1-22)式的一般均衡分析也可归入经济控制论的静态经济系统求解的范畴.

例 1 设全社会有 n 种消费品,每种消费品的生产过程中只投入两种要素:资本与劳动工时.第 i 种消费品生产函数为

$$Q_i = A_i[\delta_i L_i^{\sigma_i} + (1-\delta_i)K_i^{\sigma_i}]^{1/\sigma_i}, \tag{1-23}$$

其中,Q_i 为第 i 种产品产出量,K_i 为投入的资本,L_i 为投入的劳动工时,A_i,δ_i,σ_i 为常数,$\sigma_i < 0$.

依(1-15)式所示的利润最大法则,可求出单位产出时的劳动工时需求量 l_i 及资本需求量 k_i,它们都是劳动工时的价格 w(即工资率)及使用资本的价格 r(即租金)的函数:

$$\begin{cases} l_i = \dfrac{L_i}{Q_i} = l_i(r,w) = \dfrac{(w/\delta_i)^{1/(\sigma_i-1)}}{A_i[\delta_i(w/\delta_i)^{\sigma_i/(\sigma_i-1)} + (1-\delta_i)(r/(1-\delta_i))^{\sigma_i/(\sigma_i-1)}]^{1/\sigma_i}}; \\ k_i = \dfrac{K_i}{Q_i} = k_i(r,w) = \dfrac{(r/(1-\delta_i))^{1/(\sigma_i-1)}}{A_i[\delta_i(w/\delta_i)^{\sigma_i/(\sigma_i-1)} + (1-\delta_i)(r/(1-\delta_i))^{\sigma_i/(\sigma_i-1)}]^{1/\sigma_i}}. \end{cases} \tag{1-24}$$

依(1-21)式可求出按成本定产品价格的方程

$$p_i = wl_i(r,w) + rk_i(r,w). \tag{1-25}$$

设全社会有 h 种消费者,第 j 种消费者的效用函数为

$$U_j = B_j[a_{1j}{}^{1/\varphi_j}x_{1j}{}^{(\varphi_j-1)/\varphi_j} + \cdots + a_{nj}{}^{1/\varphi_j}x_{nj}{}^{(\varphi_j-1)/\varphi_j}]^{\varphi_j/(\varphi_j-1)}, \tag{1-26}$$

其中,x_{ij} 为第 j 种消费者对第 i 种产品的消费量.

从(1-8)式可得到第 j 种消费者的需求函数为

$$x_{ij} = \frac{a_{ij}p_i^{-\varphi_j}M_j}{a_{1j}p_1{}^{1-\varphi_j} + \cdots + a_{nj}p_n^{1-\varphi_j}}, \tag{1-27}$$

其中,M_j 为第 j 种消费者消费支出总额.当第 j 种消费者拥有 K_j^s 的资本,并付出 L_j^S 的劳动工时的情况下,其收入

$$M_j = wL_j^s + rK_j^s. \tag{1-28}$$

产品产出量等于需求量,即

$$Q_i = \sum_{j=1}^{h} x_{ij}. \tag{1-29}$$

要素需求量为

$$\begin{cases} L_i = l_iQ_i; \\ K_i = k_iQ_i. \end{cases} \tag{1-30}$$

要素需求量应等于要素供给量,即

$$\begin{cases} \sum\limits_{i=1}^{n} L_i = \sum\limits_{j=1}^{h} L_j^s; \\ \sum\limits_{i=1}^{n} K_i = \sum\limits_{j=1}^{h} K_j^s. \end{cases} \tag{1-31}$$

其中，L_j^s 及 K_j^s 为给定的已知数.

(1-24)、(1-25)、(1-27)、(1-28)、(1-29)、(1-30)、(1-31) 式构成**非线性静态供求平衡系统**. 方程数个数与未知数个数是一样的，但方程并不相互独立，因此只能求出供求平衡时市场比价 $p_1 : p_2 : \cdots : p_n : r : w$. 求出比价之后，便可确定产品供求量及投入要素的需求量等变量的数值.

例 2　设共有 n 种产品，第 i 种产品的生产函数为

$$x_i = \min\left\{\frac{x_{1i}}{a_{1i}}, \frac{x_{2i}}{a_{2i}}, \cdots, \frac{x_{ni}}{a_{ni}}, \frac{L_i}{l_i}\right\}, \quad i = 1, \cdots, n, \tag{1-32}$$

其中，x_{ji} 为生产第 i 种产品的过程中所投入的第 j 种产品的量，L_i 为投入的劳动工时量，a_{ji} 及 l_i 分别为产品投入系数及劳动工时投入系数.

如果第 i 种产品产出量为 x_i，那么投入的各种要素应成恰当比例，并且其数量为：$x_{ji} = a_{ji}x_i, j = 1, \cdots, n; L_i = l_i x_i$. 当产出向量 $\boldsymbol{x} = [x_1, \cdots, x_n]^{\mathrm{T}}$ 时，投入各产品量相应为 $\boldsymbol{Ax}$，劳动工时为 $\boldsymbol{lx}$，即

$$\boldsymbol{Ax} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{lx} = [l_1, \cdots, l_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

其中，$\boldsymbol{A}$ 称为消耗系数阵.

如果 $\boldsymbol{lx}$ 单位的劳动工时所消费的各种产品量为 $\boldsymbol{c} = [c_1, \cdots, c_n]^{\mathrm{T}}$，那么有如下的线性静态平衡方程：

$$\boldsymbol{x} = \boldsymbol{Ax} + \boldsymbol{c}, \tag{1-33}$$

上式便是**列昂惕夫（Leontief）静态投入产出平衡方程**.

如果已知消费结构 $\boldsymbol{c}$，那么相应的产品产出结构为

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{c}, \tag{1-34}$$

其中，$\boldsymbol{I}$ 为单位阵，$(\boldsymbol{I} - \boldsymbol{A})^{-1}$ 称为列昂惕夫逆阵.

1.4　广义线性多部门经济系统的运动分析

当在生产函数中考虑加工时间延迟等因素，并采用投入要素不可互相替代的生产函数时，便可得到各种类型的**线性动态多部门模型**.

1.4.1　单技术无联合产出的动态线性多部门模型

设共有 n 种产品及 1 种劳动. 每种产品只由 1 种生产过程或生产技术来生产. 第 i 种产品生产函数为

$$x_i = \min\left\{\frac{x_{1i}}{a_{1i} + b_{1i}}, \frac{x_{2i}}{a_{2i} + b_{2i}}, \cdots, \frac{x_{ni}}{a_{ni} + b_{ni}}, \frac{l_i}{l_i}\right\}, \tag{1-35}$$

其中，x_{ji} 为第 i 种产品生产过程中投入的第 j 种产品的量，投入的要素中，有些是消耗品，有些是资本品；a_{ji} 为消耗系数，表示生产 1 单位产品 i 所消耗的第 j 种产品的量；b_{ji} 为资本使用系数，表示生产 1 单位产品 i 所使用的第 j 种产品（作为资本品投入）的量. 当投入的各要素成恰当比例，即 $x_{ji} = (a_{ji} + b_{ji})x_i, j = 1, \cdots, n, L_i = l_i x_i$

时，产品产出量为 x_i.

对 n 种产品来讲，当作为消耗品投入的 n 种产品量为 $\boldsymbol{Ax}$，作为资本品投入的 n 种产品量为 $\boldsymbol{Bx}$，劳动工时投入量为 $\boldsymbol{lx}$ 时，各种产品产出量为 $\boldsymbol{x}=[x_1,\cdots,x_n]^{\mathrm{T}}$. 其中

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix},\quad \boldsymbol{B}=\begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix},\quad \boldsymbol{l}=[l_1 \quad \cdots \quad l_n].$$

1. 动态列昂惕夫投入产出模型

为了在第 $t+1$ 年产出 $\boldsymbol{x}(t+1)$，需在第 $t+1$ 年使用资本量为 $\boldsymbol{Bx}(t+1)$，在不考虑资本折旧时，要求第 t 年新增资本量为 $\boldsymbol{B}[\boldsymbol{x}(t+1)-\boldsymbol{x}(t)]$. 在第 t 年，产出 $\boldsymbol{x}(t)$ 一方面用于消耗投入 $\boldsymbol{Ax}(t)$，另一方面用于新增固定资本，余下的用于消费 $\boldsymbol{c}(t)$，因此有如下的动态平衡方程：

$$\boldsymbol{x}(t)=\boldsymbol{Ax}(t)+\boldsymbol{B}[\boldsymbol{x}(t+1)-\boldsymbol{x}(t)]+\boldsymbol{c}(t), \tag{1-36}$$

上式即为**动态列昂惕夫投入产出模型**.

2. 冯·诺伊曼(Von Neumann)生产活动分析模型

为了在第 $t+1$ 年产出 $\boldsymbol{x}(t+1)$，需在第 t 年投入消耗品及资本品的量为 $\boldsymbol{Ax}(t+1)+\boldsymbol{Bx}(t+1)$，第 $t-1$ 年投入的 $\boldsymbol{Ax}(t)+\boldsymbol{Bx}(t)$ 经 1 年使用后在第 t 年余下 $\boldsymbol{Bx}(t)$. 因此，第 t 年的产出 $\boldsymbol{x}(t)$ 加上第 $t-1$ 年投入在第 t 年的剩余 $\boldsymbol{Bx}(t)$，可用于生产投入 $(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}(t+1)$ 及消费 $\boldsymbol{c}(t)$，可得到如下平衡方程：

$$\boldsymbol{x}(t)+\boldsymbol{Bx}(t)=(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}(t+1)+\boldsymbol{c}(t), \tag{1-37}$$

上式称为**冯·诺伊曼生产活动分析模型**. (1-37) 式又可改写为

$$\boldsymbol{x}(t)=\boldsymbol{Ax}(t+1)+\boldsymbol{B}[\boldsymbol{x}(t+1)-\boldsymbol{x}(t)]+\boldsymbol{c}(t). \tag{1-38}$$

由 (1-36) 式及 (1-38) 式可知，由于 $\boldsymbol{A}$，$\boldsymbol{B}$ 阵有不同的含义，且投入到产出的时间延迟不同，则得到相应不同的平衡方程式. 类似地，可以有其它各种类型的线性多部门平衡方程式.

1.4.2 无联合生产的广义线性多部门系统的运动分析

将 (1-36) 式所示的动态投入产出模型改写为

$$\boldsymbol{Bx}(t+1)=\boldsymbol{Rx}(t)-\boldsymbol{c}(t), \tag{1-39}$$

其中 $\boldsymbol{R}=\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{B}$. 由于 n 种产品中有些产品不能作为资本品，因此 $\boldsymbol{B}$ 阵中有些行为零，即 $\boldsymbol{B}$ 阵为奇异阵. (1-39) 式所示的经济系统称为**广义动态系统**.

1. 矩阵的幂零指数

如果 $n\times n$ 方阵 $\boldsymbol{N}$ 成立，$\boldsymbol{N},\boldsymbol{N}^2,\cdots,\boldsymbol{N}^{k-1}$ 都不为零阵，而 $\boldsymbol{N}^k=\boldsymbol{0}$，则称 $\boldsymbol{N}$ 为幂零矩阵，k 称为 $\boldsymbol{N}$ 阵的幂零指数，并记为：$\mathrm{Ind}(\boldsymbol{N})=k$.

2. 性质

对任意 1 个 $n\times n$ 方阵 $\boldsymbol{A}$，必存在 1 个 $\boldsymbol{T}$ 阵将 $\boldsymbol{A}$ 化为

$$\boldsymbol{A}=\boldsymbol{T}\begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{N} \end{bmatrix}\boldsymbol{T}^{-1}, \tag{1-40}$$

其中，$\boldsymbol{G}$ 为满秩方阵，$\boldsymbol{N}$ 为幂零矩阵.

3. 矩阵的德拉金(Drazin)逆

如果 A 阵化为(1-40)式的形式,则 A 的德拉金逆

$$A^{D} = T\begin{bmatrix} G^{-1} & 0 \\ 0 & 0 \end{bmatrix} T^{-1}. \tag{1-41}$$

当 A 为满秩方阵时,$A^{D} = A^{-1}$;当 A 为幂零矩阵时,$A^{D} = 0$.

4. 广义动态投入产出系统运动分析

考虑(1-39)式所示动态投入产出系统,令

$$\hat{B} = (\lambda B + R)^{-1} B,$$

$$\hat{R} = \lambda(\lambda B + R)^{-1} R,$$

$$\hat{c}(t) = (\lambda B + R)^{-1} c(t),$$

其中,λ 是使$(\lambda B + R)^{-1}$ 存在的任意实数.那么当给定初始产出结构 $x(0)$ 及消费结构变化 $c(t)$ 时,系统的解为

$$x(t) = (\hat{B}^{D}\hat{R})^{t}\hat{B}^{D}\hat{B}x(0)(-\hat{B}^{D}\sum_{i=0}^{t-1}(\hat{B}^{D}\hat{R})^{t-i-1}\hat{c}(i) + (I - \hat{B}\hat{B}^{D})\sum_{i=0}^{k-1}(\hat{B}\hat{R}^{D})^{i}\hat{R}^{D}\hat{c}(t+i), \tag{1-42}$$

其中,$k = \operatorname{Ind}(\hat{B})$.初始值 $x(0)$ 应满足

$$x(0) = \hat{B}^{D}\hat{B}x(0) + (I - \hat{B}\hat{B}^{D})\sum_{i=0}^{k-1}(\hat{B}\hat{R}^{D})^{i}\hat{R}^{D}\hat{c}(i). \tag{1-43}$$

1.4.3 广义线性多部门经济系统运动的特征分析

当采用投入要素不可互相替代的生产函数时,产品的供给与需求动态平衡方程可化为如下形式:

$$Ex(t+1) = Hx(t) + Du(t). \tag{1-44}$$

例如,当 $E = B, H = R, D = -I$ 时,上式即为(1-39)式所示的动态投入产出系统.

1. 广义经济系统自由运动的特征分析

当(1-44)式所示系统输入向量 $u(t) = 0$ 时,(1-44)式化为

$$Ex(t+1) = Hx(t). \tag{1-45}$$

设 λ 是广义特征方程

$$|zE - H| = 0 \tag{1-46}$$

的一个根,h 是相应于 λ 的广义特征向量,

$$\lambda Eh = Hh, \tag{1-47}$$

那么 $x(t) = \lambda^{t}h$ 便是(1-45)式所示系统的一个特解.

如果(1-46)式所示广义特征方程全部非零特征根两两相异为 $\lambda_1, \cdots, \lambda_p, p \leqslant n$, n 是 $x(t)$ 的维数,它们相应的特征向量为 $h_1, \cdots, h_p$,那么如下所示函数是(1-45)式的解:

$$\begin{cases} x(t) = a_1 h_1 \lambda_1^t + \cdots + a_p h_p \lambda_p^t; \\ x(0) = a_1 h_1 + \cdots + a_p h_p. \end{cases} \tag{1-48}$$

如果(1-46)式所示**广义特征方程**为

$$|zE - H| = (z-\lambda_1)^{s_1}\cdots(z-\lambda_p)^{s_p}, s_1 + \cdots + s_p \leqslant n, \tag{1-49}$$

它的非零根为 $\lambda_1, \cdots, \lambda_p$，重数分别为 $s_1, \cdots, s_p$，那么(1-45)式的解具有如下形式：

$$x(t) = a_1^0 h_1^0 \lambda_1^t + a_1^1 h_1^1 t\lambda_1^t + \cdots + a_1^{s_1-1} h_1^{s_1-1} t^{s_1-1} \lambda_1^t + \cdots + a_p^0 h_p^0 \lambda_p^t + a_p^1 h_p^1 t\lambda_p^t + \cdots + a_p^{s_p-1} h_p^{s_p-1} t^{s_p-1} \lambda_p^t, \tag{1-50}$$

其中，a_i^j 为常数.

例 3 考虑(1-36)式的动态投入产出系统.如果消费系数阵为 T，即 $C(t) = Tx(t)$，则式(1-36)化为

$$Bx(t+1) = (I - A - T + B)x(t), \tag{1-51}$$

其中,投资系数阵 B 及消耗系数阵 A 与消费系数阵 T 的参数分别为

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 10 & 0.1 & 0 \\ 20 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 0.2 \\ 0.5 & 0 & 0 \\ 1 & 3 & 0 \end{bmatrix}, \quad T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0.02 \end{bmatrix}.$$

首先，求 $|zB - R| = 0$ 的非零根，$R = I - A - T + B$.它的两个根分别为

$$\lambda_1 = 1.025311, \lambda_2 = 23.090649.$$

其次，求相应的广义特征向量.它们分别是

$h_1 = [1, 0.755021, 5]^T$，它是相应于 λ_1 的右特征向量；

$h_2 = [1, -183.12208, 5]^T$，它是相应于 λ_2 的右特征向量.

那么，(1-51)式的解为

$$x(t) = a_1 \begin{bmatrix} 1 \\ 0.755021 \\ 5 \end{bmatrix} \times 1.025311^t + a_2 \begin{bmatrix} 1 \\ -183.12208 \\ 5 \end{bmatrix} \times 23.090649^t. \tag{1-52}$$

其中，a_1 与 a_2 的值由初始条件 $x(0)$ 决定.

由(1-52)式可知，当各部门比例恰好为：1 : 0.755021 : 5时，$a_2 = 0$，这时各部门**平衡增长速度**为 2.5311%.

2.广义经济系统强迫运动的特征分析

在(1-44)式所示的广义动态系统中，如果输入向量 $u(t)$ 的各分量是由 λ^t, $t\lambda^t$, $t^2\lambda^t$, … 形式的函数线性叠加而成，那么 $u(t)$ 可由如下形式的方程描述：

$$u(t+1) = Mu(t), \tag{1-53}$$

其中，$u(0)$ 给定，M 为常阵.

可以将(1-44)式与(1-53)式一起写成

$$\begin{bmatrix} E & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x(t+1) \\ u(t+1) \end{bmatrix} = \begin{bmatrix} H & D \\ 0 & M \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}. \tag{1-54}$$

由上式可知，这时(1-44)式广义系统强迫运动分析问题便化为自由运动分析问题.

1.5 非线性多部门经济系统的运动分析

当采用投入要素可以相互替代的生产函数时,讨论市场供求平衡将得到非线性动态多部门经济系统.对这一类型的动态系统来讲,要用具体的数学表达式表示产出量随时间变化的全过程是一件十分困难的事情.在实际应用中更感兴趣的是讨论系统的平衡增长轨道及最优增长轨道.

2 经济系统运动的平衡增长轨道

经济系统运动的平衡增长轨道是一条供求平衡的协调增长轨道.

2.1 动态投入产出系统的平衡增长

2.1.1 已知消费结构时动态投入产出系统平衡增长率与产出结构的计算

考虑(1-35)式所示的生产函数,当第 t 年产出为 $\boldsymbol{x}(t)$ 时,劳动工时投入量为 $l_1x_1+\cdots+l_nx_n=\boldsymbol{lx}(t)$,其中 $\boldsymbol{l}=[l_1,\cdots,l_n]$.如果每单位劳动工时消费的各种产品量为$[d_1,\cdots,d_n]^{\mathrm{T}}$,那么 $\boldsymbol{lx}$ 单位劳动工时消费各种产品量为

$$\boldsymbol{c}(t)=\begin{bmatrix}d_1\\ \vdots\\ d_n\end{bmatrix}[l_1,\cdots,l_n]\begin{bmatrix}x_1(t)\\ \vdots\\ x_n(t)\end{bmatrix}=\begin{bmatrix}d_1l_1\cdots d_1l_n\\ \vdots\\ d_nl_1\cdots d_nl_n\end{bmatrix}\begin{bmatrix}x_1(t)\\ \vdots\\ x_n(t)\end{bmatrix}=\boldsymbol{Tx}(t),\quad(2\text{-}1)$$

其中,$\boldsymbol{T}$ 为消费系数阵.将它代入(1-36)式,可得到在给定消费结构下的动态投入产出模型

$$\boldsymbol{Bx}(t+1)=(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T}+\boldsymbol{B})\boldsymbol{x}(t).\quad(2\text{-}2)$$

当平衡增长时,各部门有相同的增长率 λ,即

$$\boldsymbol{x}(t+1)=(1+\lambda)\boldsymbol{x}(t).\quad(2\text{-}3)$$

将它代入(2-2)式,可知平衡增长率 λ 是如下广义特征方程的解:

$$|(1+\lambda)\boldsymbol{B}-(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T}+\boldsymbol{B})|=0.\quad(2\text{-}4)$$

求解平衡增长轨道要点如下:

(1) 对任何一个非负实方阵,存在一个模最大的非负特征根,称之为庇隆-弗罗宾纽斯(Perron-Frobenius)根以及相应的非负特征向量.

(2) 当非负阵$(\boldsymbol{A}+\boldsymbol{T})$满足郝庆芝-西蒙(Hawkins-Simon)条件,即它的最大模特征根小于1时,$(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}$ 也为非负阵,因此$(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}\boldsymbol{B}$ 也为非负阵.

(3)(2-4)式特征方程可化为

$$|(1/\lambda)\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}\boldsymbol{B}|=0.\quad(2\text{-}5)$$

从上式可看出平衡增长率 λ 是非负阵$(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}\boldsymbol{B}$ 的庇隆-弗罗宾纽斯根的倒

数,产出比例或产出结构为相应的非负特征向量 x^*.

$$(1/\lambda)x^* = (I - A - T)^{-1}Bx^*. \tag{2-6}$$

2.1.2 已知消费结构时动态投入产出系统平衡增长的价格与利润率的计算

设平衡增长时各种产品价格为 $p = [p_1, \cdots, p_n]$,在第 t 年产出为 $x(t)$,收入为 $px(t)$. 为了产出 $x(t)$ 所消耗的各种产品量为 $Ax(t)$,其价值为 $pAx(t)$. 同时,为产出 $x(t)$,需在第 t 年使用固定资本 $Bx(t)$,其价值为 $pBx(t)$. 设利息率为 β,那么使用这些固定资本所付利息为 $\beta pBx(t)$. 此外,为产出 $x(t)$ 需投入劳动工时 $lx(t)$,所支付的工资额为 $wlx(t)$,w 为工资率. 当系统平衡增长时,收入等于支出,因此有如下平衡方程:

$$px(t) = pAx(t) + \beta pBx(t) + wlx(t),$$

或

$$p = wl(I - A - \beta B)^{-1}. \tag{2-7}$$

从上式可以看出,β 既可以看做平衡增长时的利息率,又可看做资本利润率. 因为利润等于收入 $px(t)$ 减去成本 $pAx(t) + wlx(t)$. 将利润除以使用的资本 $pBx(t)$ 即为资本利润率.

当给定工资率 w 之后,市场均衡价格由(2-7)式求解.

w 是每单位劳动工时的工资,如果在市场价格之下购买各种产品量是已知的,即已知单位劳动工时消费结构为 d,则有

$$w = [p_1, \cdots, p_n]\begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = pd.$$

将它代入(2-7)式,得

$$p = pdl(I - A - \beta B)^{-1} = pT(I - A - \beta B)^{-1},$$

注意到 $dl = T$ 为消费系数阵,上式又可记为

$$(1/\beta)p = pB(I - A - T)^{-1}. \tag{2-8}$$

从上式可知,协调发展利润率 β 是非负阵 $B(I - A - T)^{-1}$ 的庇隆 - 弗罗宾纽斯根的倒数,产品比价是相应的左特征行向量.

由于 $(I - A - T)^{-1}B$ 与 $B(I - A - T)^{-1}$ 的特征根相同,从(2-6)式与(2-8)式可知,平衡增长利润率 β 与增长率 λ 是相等的.

2.1.3 未知消费结构时动态投入产出系统平衡增长率、产出结构、价格、利润率的计算

如果每单位劳动工时的工资 w 所购买各种产品量 d 是市场价格 p 及工资率 w 的函数,即

$$d = f(p, w), \tag{2-9}$$

上式是在预算约束 $pd = w$ 之下求效用最大所得到的需求函数,那么市场平衡增长时价格 p、平衡增长率 λ、产出结构 x、利润率 β 可由(2-6)式、(2-7)式、(2-9)式求解,即

$$\begin{cases} d = f(p, w); \\ (1/\lambda)x = (I - A - T)^{-1}Bx, T = dl; \\ p = wl(I - A - \beta B)^{-1}. \end{cases} \tag{2-10}$$

例1 考虑如下动态投入产出系统：

$$x(t) = Ax(t) + B[x(t+1) - x(t)] + c(t). \tag{2-11}$$

其参数为

$$A = \begin{bmatrix} 0.6 & 0.3 & 0 \\ 0 & 0 & 0.5 \\ 0.2 & 0.3 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 10 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad l = [0.5, 0.5, 0.3].$$

(2-9) 式所示的需求函数为如下的线性支出系统：

$$\begin{cases} p_1 d_1 = 0.012p_1 - 0.01p_2 - 0.004p_3 + 0.2w; \\ p_2 d_2 = -0.0075p_1 + 0.025p_2 - 0.01p_3 + 0.5w; \\ p_3 d_3 = -0.0045p_1 - 0.015p_2 + 0.014p_3 + 0.3w. \end{cases} \tag{2-12}$$

求解上述模型与参数下平衡增长率、利润率、产出结构、价格的步骤如下：

第1步：求动态投入产出系统所能达到的最大增长率 λ^*，即令消费阵 $T = 0$，由(2-6)式可知，$(1/\lambda^*)x = (I - A)^{-1}Bx$，求出 $\lambda^* = 0.31$.

第2步：确定实际增长率 λ，$0 < \lambda < \lambda^*$. 在本例中，取 $\lambda = 0.1 = 10\%$.

第3步：计算平衡增长时的价格. 由于平衡增长时只能求出比价，因此可令 $w = 1$. 此外，由于平衡增长率 λ 等于利润率 β，故 $\beta = 0.1$，将它代入(2-7)式，求出价格为：$p_1 = 2.547619$，$p_2 = 4.590476$，$p_3 = 2.59238$.

第4步：计算消费结构. 将求出的价格代入(2-12)式，可直接求出消费结构：$d_1 = 0.0684112$，$d_2 = 0.1241053$，$d_3 = 0.0986468$.

第5步：计算产出结构. 当求出消费结构 $d = [d_1, d_2, d_3]^T$ 后，便可求出消费系数阵 $T = dl$，再将它代入(2-6)式便可求出产出结构：$x_1 : x_2 : x_3 = 3.7212 : 1 : 1.316$.

2.2 冯·诺伊曼生产活动分析模型的平衡增长

一般地说，单技术无联合产出线性多部门模型可以用(1-44)式所示广义线性方程来描述. 例如，(1-38)式的冯·诺伊曼生产活动分析模型可化为(1-44)式的形式，这时 $E = A + B$，$H = I + B$，$D = -I$，$u(t) = c(t)$. 冯·诺伊曼生产活动分析模型的平衡增长轨道计算要点如下：

(1) 求出产出与消费同步平衡增长的条件.

将(1-38)式的线性方程记为

$$(A + B)x(t+1) = (I + B)x(t) - c(t). \tag{2-13}$$

如果产出 $x(t)$ 与消费 $c(t)$ 按同一增长率 λ 增长，即

$$\begin{cases} x(t) = x(0)(1+\lambda)^t; \\ c(t) = c(0)(1+\lambda)^t, \end{cases}$$

将它代入(2-13)式中,可求出初始产出结构 $\boldsymbol{x}(0)$ 与消费结构 $\boldsymbol{c}(0)$ 所应满足的同步增长条件,即

$$\boldsymbol{x}(0)=[\boldsymbol{I}-\boldsymbol{A}-\lambda(\boldsymbol{A}+\boldsymbol{B})]^{-1}\boldsymbol{c}(0). \tag{2-14}$$

(2) 如果已知每单位劳动工时对各种产品的消费结构为 $\boldsymbol{d}=[d_1,\cdots,d_n]^{\mathrm{T}}$,那么为了在第 $t+1$ 年产出 $\boldsymbol{x}(t+1)$,需在第 t 年投入劳动工时 $\boldsymbol{lx}(t+1)$,其相应消费量为 $\boldsymbol{dlx}(t+1)=\boldsymbol{Tx}(t+1)$,$\boldsymbol{T}$ 为消费系数阵.将第 t 年消费 $\boldsymbol{c}(t)=\boldsymbol{Tx}(t+1)$ 代入(2-13)式,得到

$$(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}(t+1)=(\boldsymbol{I}+\boldsymbol{B})\boldsymbol{x}(t)-\boldsymbol{Tx}(t+1),$$

在平衡增长时,$\boldsymbol{x}(t+1)=(1+\lambda)\boldsymbol{x}(t)$,代入上式得到

$$(1/\lambda)\boldsymbol{x}(t)=(\boldsymbol{I}-\boldsymbol{A}-t)^{-1}(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{T})\boldsymbol{x}(t), \tag{2-15}$$

从上式可以看出,当$(\boldsymbol{A}+\boldsymbol{T})$阵满足郝庆芝 - 西蒙条件时,平衡增长率 λ 是非负阵 $(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{T})$ 的庇隆 - 弗罗宾纽斯根的倒数,产出结构为相应的右特征列向量.

(3) 分析平衡增长轨道的存在性与唯一性.

如果$(\boldsymbol{A}+\boldsymbol{T})$阵满足郝庆芝 - 西蒙条件,且非负方阵$(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{T})$为不可分解矩阵,那么该阵存在唯一的最大模正特征根$(1/\lambda^*)$与相应的正右特征产出向量 $\boldsymbol{x}^*$.不可分解经济系统的含义是:任何一个经济部门都直接或间接地与其它所有部门发生投入产出关系.

(4) 如果每单位劳动工时所得的工资 w 在市场价格下购买各种产品量由需求函数 $\boldsymbol{d}=\boldsymbol{f}(\boldsymbol{p},w)$ 来表示,那么在平衡增长轨道上产出增长率 λ、产出结构 $\boldsymbol{x}$、利润率 β、物价 $\boldsymbol{p}$ 由如下方程求解:

$$\begin{cases}\boldsymbol{d}=\boldsymbol{f}(\boldsymbol{p},w);\\(1/\lambda)\boldsymbol{x}=(\boldsymbol{I}-\boldsymbol{A}-\boldsymbol{T})^{-1}(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{T})\boldsymbol{x},\ \boldsymbol{T}=\boldsymbol{dl};\\\boldsymbol{p}=(1+\beta)w\boldsymbol{l}[\boldsymbol{I}-\boldsymbol{A}-\beta(\boldsymbol{A}+\boldsymbol{B})]^{-1}.\end{cases} \tag{2-16}$$

上式与(2-10)式求解方法类似.

其他类型的单技术无联合产出线性多部门模型的平衡增长轨道的计算方法可作类似讨论.

2.3 非线性多部门动态经济系统的平衡增长

以上各节所讨论的线性多部门经济系统,采用的都是投入要素不可互相替代的生产函数.例如,采用(1-32)式所示的生产函数,将得到(1-33)式所示的列昂惕夫静态投入产出模型.如果在(1-32)中考虑加工时间,则有

$$x_i(t+1)=\min\left\{\frac{x_{1i}(t)}{a_{1i}},\frac{x_{2i}(t)}{a_{2i}},\cdots,\frac{x_{ni}(t)}{a_{ni}},\frac{l_i(t)}{l_i}\right\}, \tag{2-17}$$

其中,$i=1,\cdots,n$,$x_{ji}(t)$ 表示在第 t 年为生产第 i 种产品所投入的第 j 种产品的量,$L_i(t)$ 为第 t 年第 i 种产品生产过程中投入的劳动工时量.

与(1-33)式讨论类似,基于(2-17)式的生产函数可得到如下的冯·诺伊曼 - 列昂惕夫模型:

$$x(t)=Ax(t+1)+c(t). \tag{2-18}$$

类似地，如果在(1-35)式中考虑投入到产出之间的时间延迟，那么可得到(1-36)式及(1-37)式所示的列昂惕夫动态投入产出模型及冯·诺伊曼生产活动分析模型.

自20世纪60年代以来，许多经济学家力图将线性多部门模型推广到非线性多部门模型的情况. 例如，20世纪80年代 Takao Fujimoto 等人将(1-33)式静态投入产出模型推广而得到如下的**非线性静态投入产出模型**:

$$x=A(x)\cdot x+c, \tag{2-19}$$

其中，$A(x)$ 为非线性投入系数阵，其元素是 x 的非线性多元函数.

(2-19)式仅是(1-33)式静态线性投入产出模型在形式上的推广，其中投入系数阵 $A(x)$ 的参数难以在实际中获得.

下面分析如何从生产函数上进行从线性到非线性的推广. 首先将投入要素不可替代且考虑加工时间的生产函数(2-17)式推广为投入要素可互相替代且考虑加工时间的生产函数:

$$x_i(t+1)=A_i\Big[\sum_{j=1}^{n}a_{ji}x_{ji}^{\sigma_i}(t)+l_iL_i^{\sigma_i}(t)\Big]^{1/\sigma_i}, \tag{2-20}$$

其中，$i=1,\cdots,n$，$-\infty<\sigma_i<1$，$A_i>0$，$x_{ji}(t)$ 是第 t 年为生产第 i 种产品所投入的第 j 种产品的量，$L_i(t)$ 为劳动工时投入量，$x_i(t+1)$ 为第 $t+1$ 年第 i 种产品产出量.

对(2-20)式所示非线性动态多部门经济系统来讲，其平衡增长轨道上的产出结构、增长率、利润率、物价等都有相应计算公式. 与线性多部门列昂惕夫模型及冯·诺伊曼模型等相对应有一系列计算公式及计算方法. 例如，(2-20)式所示系统在平衡增长轨道上的产品价格 $p=[p_1,\cdots,p_n]$、利润率 β、工资率 w 由下式计算:

$$p=(1+\beta)p\,A(p,w)+(1+\beta)w\,\Phi(p,w), \tag{2-21}$$

其中，$A(p,w)$ 为 $n\times n$ 矩阵:

$$A(p,w)=\begin{bmatrix}A_{11}(p,w) & \cdots & A_{1n}(p,w)\\ \vdots & & \vdots\\ A_{n1}(p,w) & \cdots & A_{nn}(p,w)\end{bmatrix},$$

其元素为

$$A_{ij}(p,w)=\frac{p_i^{1/(\sigma_j-1)}a_{ij}^{1/(1-\sigma_j)}}{A_j\Big[\sum_{i=1}^{n}a_{ij}^{1/(1-\sigma_j)}p_i^{\sigma_j/(\sigma_j-1)}+l_j^{1/(1-\sigma_j)}w^{\sigma_j/(\sigma_j-1)}\Big]^{1/\sigma_j}}.$$

在(2-21)式中，$\Phi(p,w)=[\Phi_1(p,w),\cdots,\Phi_n(p,w)]$，其元素为

$$\Phi_j(p,w)=\frac{w^{1/(\sigma_j-1)}l_j^{1/(1-\sigma_j)}}{A_j\Big[\sum_{i=1}^{n}a_{ij}^{1/(1-\sigma_j)}p_i^{\sigma_j/(\sigma_j-1)}+l_j^{1/(1-\sigma_j)}w^{\sigma_j/(\sigma_j-1)}\Big]^{1/\sigma_j}}.$$

由于经济系统只有比价才有意义，因此在(2-20)式中可设 $w=1$，那么一旦选择一个可行的利润率 β，便可求出相应的价格 p.

实际应用中，可依计量经济学知识及历史数据估计系统参数，按简易的计算方法求出平衡增长解.

3 市场调节的稳定性分析

3.1 产品市场调节的稳定性分析

设有 n 种产品，第 i 种产品的供给量 S_i 及需求量 D_i 是价格 $p_1,\cdots,p_n$ 的函数（其它条件不变）. 价格由市场供求决定. 当供大于求时，价格将下降；反之，当供不应求时，价格上升. 可以用各种数学公式来近似描述这种价格变化过程. 例如，如果第 i 种产品价格单位时间内上升或下降的百分比与供求差额上升或下降的百分比成正比，那么价格变化可由下式描述：

$$\frac{\mathrm{d}p_i/\mathrm{d}t}{p_i} = k_i\frac{D_i(p_1,\cdots,p_n) - S_i(p_1,\cdots,p_n)}{D_i(p_1,\cdots,p_n)}, \tag{3-1}$$

其中，$k_i > 0, i = 1,\cdots,n$.

如果价格变化仅与供求差额 E_i 成正比，那么价格变化可由下式描述：

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = k_iE_i = k_i[D_i(p_1,\cdots,p_n) - S_i(p_1,\cdots,p_n)], \tag{3-2}$$

其中，$k_i > 0, i = 1,\cdots,n$.

产品市场调节的稳定性问题是：当市场价格变化由(3-1) 式或(3-2) 式等描述时，在时间 t 足够大之后能否使市场价格自动趋于供求平衡价格，即

$$\lim_{t\to\infty} p_i(t) = p_i^* \text{ 是否成立},$$

其中，p_i^* 是市场均衡价格，它使得

$$D_i(p_1^*,\cdots,p_n^*) = S_i(p_1^*,\cdots,p_n^*), \quad i = 1,\cdots,n.$$

可用**李雅普诺夫能量函数**讨论(3-1) 式或(3-2) 式所示非线性动态经济系统的稳定性问题.

首先，构造如下的李雅普诺夫能量函数 $V(t)$：

$$V(t) = \frac{1}{2}\sum_{i=1}^{n} k_iE_i^2, \tag{3-3}$$

其中，$E_i = D_i - S_i, i = 1,\cdots,n$. 从上式可知，$V(t) \geqslant 0$，仅在平衡点为零，即仅当 $p_i = p_i^*, i = 1,\cdots,n$ 时，$V(t) = 0$.

然后，让(3-3) 式中的能量函数 $V(t)$ 对时间 t 求导，得到

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = [k_1E_1,\cdots,k_nE_n]\begin{bmatrix} \frac{\partial E_1}{\partial p_1} & \cdots & \frac{\partial E_1}{\partial p_n} \\ \vdots & & \vdots \\ \frac{\partial E_n}{\partial p_1} & \cdots & \frac{\partial E_n}{\partial p_n} \end{bmatrix}\begin{bmatrix} \frac{\mathrm{d}p_1}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}p_n}{\mathrm{d}t} \end{bmatrix}. \tag{3-4}$$

如果采用(3-1) 式的定价策略，将(3-1) 式代入上式求得

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t}=[k_1E_1,\cdots,k_nE_n]\begin{bmatrix}\eta_{11} & \cdots & \eta_{1n}\\ \vdots & & \vdots\\ \eta_{n1} & \cdots & \eta_{nn}\end{bmatrix}\begin{bmatrix}k_1E_1\\ \vdots\\ k_nE_n\end{bmatrix}, \tag{3-5}$$

其中,$\eta_{ij}=(\partial E_i/\partial p_j)\times p_j/D_i$.

如果采用(3-2)式的定价策略,将(3-2)式代入(3-4)式求得

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t}=[k_1E_1,\cdots,k_nE_n]\begin{bmatrix}\frac{\partial E_1}{\partial p_1} & \cdots & \frac{\partial E_1}{\partial p_n}\\ \vdots & & \vdots\\ \frac{\partial E_n}{\partial p_1} & \cdots & \frac{\partial E_n}{\partial p_n}\end{bmatrix}\begin{bmatrix}k_1E_1\\ \vdots\\ k_nE_n\end{bmatrix}. \tag{3-6}$$

一般地说,(3-5)式及(3-6)式中矩阵主对角线上元素值为负.即第i种产品价格上升,将引起该种产品需求量下降、供给量上升,也就是E_i下降;反之,p_i下降将引起E_i上升.因此,$\partial E_i/\partial p_i<0$.由于一种产品价格变化主要影响该种产品的供求变化,因此(3-5)式及(3-6)式中矩阵主对角线上元素在绝对值上占主导地位.这意味着$\mathrm{d}V(t)/\mathrm{d}t\leqslant 0$.当然,这只是在经验上推测,在一般情况下$\mathrm{d}V(t)/\mathrm{d}t$将小于或等于零.如果对于供求函数的具体表达式能证明能量函数对时间t求导$\mathrm{d}V(t)/\mathrm{d}t\leqslant 0$,且仅在平衡点为零,这意味着系统能量将不断减少,直至为零,同时也意味着系统终将到达平衡点,从而系统是渐近稳定的.

应当注意到$V(t)\geqslant 0$,$\mathrm{d}V(t)\leqslant 0$仅是系统渐近稳定的必要条件而不是充分条件.

例1 设共有2种产品,供给量不变,为常数S_1,S_2;需求函数如(1-8)式所示($\sigma=2$),即

$$\begin{cases}D_1=\dfrac{a_1p_1^{-2}M}{a_1p_1^{-1}+a_2p_2^{-1}},\\ D_2=\dfrac{a_2p_2^{-2}M}{a_1p_1^{-1}+a_2p_2^{-1}}.\end{cases}$$

市场定价策略由(3-2)式描述($i=2$).

按(3-3)式构造能量函数$V(t)\geqslant 0$,且仅在平衡点为零:

$$\begin{aligned}V(t)&=\frac{k_1}{2}\left(\frac{a_1p_1^{-2}M}{a_1p_1^{-1}+a_2p_2^{-1}}-S_1\right)^2+\frac{k_2}{2}\left(\frac{a_2p_2^{-2}M}{a_1p_1^{-1}+a_2p_2^{-1}}-S_2\right)^2\\&=k_1E_1^2/2+k_2E_2^2/2.\end{aligned}$$

让能量函数$V(t)$对t求导,得

$$\begin{aligned}\frac{\mathrm{d}V}{\mathrm{d}t}&=\frac{-k_1^2E_1^2(a_1^2p_1^{-4}+2a_1a_2p_1^{-3}p_2^{-1})-k_2^2E_2^2(a_2^2p_2^{-4}+2a_1a_2p_1^{-1}p_2^{-3})}{(a_1p_1^{-1}+a_2p_2^{-1})^2}M+\\&\quad\frac{2k_1k_2E_1E_2a_1a_2p_1^{-2}p_2^{-2}}{(a_1p_1^{-1}+a_2p_2^{-1})^2}M\\&\leqslant\frac{-(k_1E_1a_1p_1^{-2}+k_2E_2a_2p_2^{-2})^2}{(a_1p_1^{-1}+a_2p_2^{-1})^2}M\leqslant 0,\end{aligned}$$

且仅在平衡点 $E_1 = E_2 = 0$时，$dV(t)/dt = 0$，因此系统是渐近稳定的，即 $t \to \infty$ 时将到达供求平衡点.

3.2 市场调节理论与鲁棒调节理论的关系

鲁棒(robust) **调节理论**是控制理论的重要分支.对线性连续时间或离散时间定常系统来讲，其鲁棒调节理论有十分丰富的内容.掌握鲁棒调节理论不仅可以深刻理解市场调节规律，而且可以用之设计各种宏观经济政策.

3.2.1 非线性静态系统目标为常向量时的鲁棒调节

考虑如下非线性静态系统：

$$\boldsymbol{y} = f(\boldsymbol{u}), \tag{3-7}$$

其中，$\boldsymbol{y}$ 是 n 维目标变量，$\boldsymbol{u}$ 是 n 维控制变量.如果希望目标变量为给定的常向量 $\boldsymbol{y}^*$，即希望偏差

$$e_i = y_i - y_i^*, \quad i = 1,\cdots,n \tag{3-8}$$

为零，那么控制策略 $u_i(t)$ 可按如下式子设计：

$$\frac{du_i}{dt} = k_i e_i, k_i > 0, \quad i = 1,\cdots,n,$$

或

$$\frac{d\boldsymbol{u}}{dt} = \boldsymbol{K}\boldsymbol{e}, \tag{3-9}$$

其中，$\boldsymbol{K}$ 为对角阵，对角线上元素依次为 $k_1,\cdots,k_n$，$\boldsymbol{e} = [e_1,\cdots,e_n]^{\mathrm{T}}$.

(3-9) 式所示系统称为**鲁棒调节器**，它的极点为零.当希望目标值 $\boldsymbol{y}^*$ 为常向量时，它是极点为零的系统的解，这时称 $\boldsymbol{y}^*$ 的极点为零.鲁棒调节器的极点与希望目标值的极点要一致，这便是所谓的"内模原理".

(3-7) ~ (3-9) 式构成如下的闭环动态系统：

$$\frac{d\boldsymbol{u}}{dt} = \boldsymbol{K}[f(\boldsymbol{u}) - \boldsymbol{y}^*]. \tag{3-10}$$

如果(3-10) 式所示系统是渐近稳定的，即成立

$$\lim_{t\to\infty} \boldsymbol{u}(t) = \text{常向量 } \boldsymbol{u}^*,$$

那么有

$$\lim_{t\to\infty} \frac{d\boldsymbol{u}}{dt} = \boldsymbol{0} = \boldsymbol{K}\boldsymbol{e},$$

这意味着目标值 $\boldsymbol{y}$ 终将达到给定值$\boldsymbol{y}^*$.

当(3-7) 式所示受控系统满足一定条件时，(3-10) 式的闭环动态系统将是渐近稳定的.

考虑矩阵

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{u}}=\begin{bmatrix}\frac{\partial y_1}{\partial u_1} & \cdots & \frac{\partial y_1}{\partial u_n}\\ \vdots & & \vdots\\ \frac{\partial y_n}{\partial u_1} & \cdots & \frac{\partial y_n}{\partial u_n}\end{bmatrix}, \tag{3-11}$$

如果该阵为准对角优势阵，即存在常数 $C_1,\cdots,C_n$ 使得

$$C_i\left|\frac{\partial y_i}{\partial u_i}\right|>\sum_{r\neq i}C_r\left|\frac{\partial y_i}{\partial u_r}\right|,\quad i=1,\cdots,n,r=1,\cdots,n \tag{3-12}$$

成立，且对角线上元为负值，那么(3-10)式所示系统是渐近稳定的. 这可以通过构造能量函数

$$V(t)=\max_i\frac{|k_ie_i|}{C_i}\geqslant 0 \tag{3-13}$$

来证明，上式仅在平衡点时 $V(t)=0$. 将 $V(t)$ 对 t 求导，可以证明 $\mathrm{d}V(t)\leqslant 0$，且仅在平衡点处为零. 因此当 t 足够大时，$V(t)$ 趋于零，从而 e_i 趋于零，$i=1,\cdots,n$.

3.2.2 市场调节与鲁棒调节的关系

与(3-7)式对应，考虑供求系统

$$\boldsymbol{e}=\boldsymbol{e}(\boldsymbol{p}). \tag{3-14}$$

其中，$\boldsymbol{e}=[e_1,\cdots,e_n]^{\mathrm{T}}$ 为超额需求向量，即第 i 种产品超额需求量 e_i 等于该种产品需求量 D_i 减去供给量 S_i；$\boldsymbol{p}=[p_1,\cdots,p_n]$ 为价格向量. 市场依供需情况不断改变价格，价格 $\boldsymbol{p}$ 的变化由下式描述：

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}=\boldsymbol{K}\boldsymbol{e}. \tag{3-15}$$

上式与(3-9)式对应，$\boldsymbol{K}$ 也是正对角阵. 市场调节的目标是达到供求平衡，即 $\boldsymbol{e}=\boldsymbol{0}$.

与(3-11)式或(3-6)式对应，考虑如下矩阵：

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{p}}=\begin{bmatrix}\frac{\partial e_1}{\partial p_1} & \cdots & \frac{\partial e_1}{\partial p_n}\\ \vdots & & \vdots\\ \frac{\partial e_n}{\partial p_1} & \cdots & \frac{\partial e_n}{\partial p_n}\end{bmatrix}. \tag{3-16}$$

如果该阵为准对角优势阵，且对角线上元素为负，即存在常数 $C_1,\cdots,C_n$，使得

$$C_i\left|\frac{\partial e_i}{\partial p_i}\right|>\sum_{r\neq i}C_r\left|\frac{\partial e_i}{\partial p_r}\right|,\quad i=1,\cdots,n,r=1,\cdots,n \tag{3-17}$$

成立，那么由(3-15)式所描述的市场调节系统将是渐近稳定的，即当时间 t 足够大时，市场达供求平衡.

(3-15)式是描述市场依供求定价的行为方程，称之为市场调节器，它与(3-9)式的鲁棒调节器相对应.

亚当·斯密在200年前就提出著名的"看不见的手"的理论，他认为每日每时有许多人生产各种产品，又有许多人购买这些产品，却看不到具体的人在管理市场，

市场之所以能自动达到供求平衡,是由于“看不见的手”在操纵市场的结果.从以上分析可以看出,“看不见的手”即为(3-15)式的市场调节器.市场能自动达到供求平衡是由于供求函数或超额需求函数满足(3-16)式的条件.(3-16)式的条件容易从直觉上予以理解,也可以给出具体的供求函数加以验证.

4 经济系统的目标设定

一般地说,经济系统的目标境界是要达到既有效益又公平.效益指的是经济系统能长期可持续快速增长且具有良好的生态环境等各项指标.公平则是**价值判断准则**,不同的人有不同的公平准则.社会主义市场经济认为按劳分配是公平的原则和境界.

4.1 生产要素与消费品配置的帕雷托最优境界

4.1.1 生产要素在生产者中的帕雷托最优配置

设有 m 种要素 $K_1,\cdots,K_m$ 分配给 l 个生产者,每个生产者的生产函数为

$$Q_i = Q_i(K_{i1},\cdots,K_{im}), \quad i = 1,\cdots,l. \tag{4-1}$$

其中,K_{ij} 为第 i 个生产者分到的第 j 种要素的量,即

$$\sum_{i=1}^{l} K_{ij} = K_j, \quad j = 1,\cdots,m. \tag{4-2}$$

设(4-1)式是投入要素可互相替代的生产函数,且任一种要素投入的增加都将引起产出的增加,即

$$\partial Q_i/\partial K_{ij} > 0, j = 1,\cdots,m.$$

1. 边际技术替代率

如果第 i 个生产者分到的各种要素量为 $K_{i1},\cdots,K_{im}$,当取走第 j 种要素 ΔK_{ij} 再给它第 k 种要素的量为 ΔK_{ik} 时,产出量 Q_i 不变,则称 $-\Delta K_{ij}/\Delta K_{ik}$ 为给定要素配置状态下第 j 种要素与第 k 种要素的边际替代率($\Delta \to 0$).由于 $\Delta Q_i = 0$,令 ΔK_{ij} 与 ΔK_{ik} 趋于零时,得

$$-\frac{\mathrm{d}K_{ij}}{\mathrm{d}K_{ik}} = \frac{\partial Q_i/\partial K_{ik}}{\partial Q_i/\partial K_{ij}}, \quad i = 1,\cdots,l. \tag{4-3}$$

2. 生产要素的帕雷托最优配置

把 m 种生产要素分配给 l 个生产者,可以有许多方案,对其中某一方案来讲,如果再也找不到新的方案使至少有一个生产者产量增加,其余生产者产量不变,则称原有分配方案为**帕雷托最优方案**.

有限的要素在生产者中进行分配应当达到帕雷托最优境界,因为不然的话总可以找到另一方案使其中某些产品产量增加,而其余产品产量不变.

有限的要素在生产者中分配达帕雷托最优的必要条件是：各种产品的生产时投入的第 j 种要素与第 k 种要素的边际替代率相等，即

$$\frac{\partial Q_i/\partial K_{ik}}{\partial Q_i/\partial K_{ij}} = \frac{\partial Q_h/\partial K_{hk}}{\partial Q_h/\partial K_{hj}}, \quad h = 1, \cdots, m. \tag{4-4}$$

3. 完善的市场机制可达到要素在生产者中的帕雷托最优配置

第 i 个生产者在给定的要素市场价格下追求利润最大，依(1-15)式的利润最大法则有

$$\frac{p_i(\partial Q_i/\partial K_{ik})}{p_i(\partial Q_i/\partial K_{ij})} = \frac{w_k}{w_j},$$

即

$$\frac{\partial Q_i/\partial K_{ik}}{\partial Q_i/\partial K_{ij}} = \frac{\partial Q_h/\partial K_{hk}}{\partial Q_h/\partial K_{hj}} = \frac{w_k}{w_j}, \tag{4-5}$$

其中，$i = 1, \cdots, l, j = 1, \cdots, m$. 上式意味着在追求利润最大及完善的市场机制下可达要素在生产者中的帕雷托最优配置.

4. 生产要素帕雷托最优配置的非唯一性

帕雷托最优境界不是唯一的，一般存在无穷多个帕雷托最优点.

4.1.2　消费品在消费者中的帕雷托最优配置

设有 n 种消费品的数量分别为 $S_1, \cdots, S_n$，将它们分配给 q 个消费者. 每个消费者效用函数为

$$U_i = U_i(D_{i1}, \cdots, D_{in}), \quad i = 1, \cdots, q. \tag{4-6}$$

其中，D_{ij} 为第 i 个消费者分到的第 j 种消费品的量，即有

$$\sum_{i=1}^{q} D_{ij} = D_j, \quad j = 1, \cdots, n.$$

1. 消费品边际替代率

如果第 i 个消费者分到的各种消费品量为 $D_{i1}, \cdots, D_{in}$，当取走第 j 种消费品 ΔD_{ij} 再给他第 k 种消费品的量为 ΔK_{ik} 时，效用不变，则称 $-\Delta D_{ij}/\Delta D_{ik}$ 为给定消费品配置状态下第 j 种消费品与第 k 种消费品的边际替代率. 由于 $\Delta U_j = 0$，令 ΔD_{ij} 与 ΔD_{ik} 为无穷小量时，得

$$-\frac{\mathrm{d}D_{ij}}{\mathrm{d}D_{ik}} = \frac{\partial U_i/\partial D_{ik}}{\partial U_i/\partial D_{ij}}, \quad i = 1, \cdots, q. \tag{4-7}$$

2. 消费品的帕雷托最优配置

把 n 种消费品分配给 q 个消费者时，可以有许多方案. 对其中某一方案来讲，如果再也找不到新的方案使至少有一个消费者效用增加，其余消费者效用不变，则称原有分配方案为帕雷托最优方案.

有限的消费品在消费者中进行分配应当达帕雷托最优境界，因为不然的话总可以找到另一方案使其中某些消费者效用增加，而其余消费者效用不变.

有限的消费品在消费者中分配达帕雷托最优的必要条件是：任意两个消费者

对任意两种消费品的边际替代率相等，即

$$\frac{\partial U_i/\partial D_{ij}}{\partial U_i/\partial D_{ik}} = \frac{\partial U_h/\partial D_{hj}}{\partial U_h/\partial D_{hk}}, \quad i = 1,\cdots,q, \quad k = 1,\cdots,n. \tag{4-8}$$

3. 完善的市场机制可达消费品在消费者中的帕雷托最优配置

在市场机制下，n 种消费品价格分别为 $p_1,\cdots,p_n$. 各个消费者在其预算约束下谋求效用最大，应满足如下的效用最大法则：

$$\frac{\partial U_i/\partial D_{i1}}{p_1} = \frac{\partial U_i/\partial D_{i2}}{p_2} = \cdots = \frac{\partial U_i/\partial D_{in}}{p_n}. \tag{4-9}$$

因此，任意两个消费者对任意两种消费品的边际替代率等于这两种消费品价格比.

$$\frac{\partial U_i/\partial D_{ij}}{\partial U_i/\partial D_{ik}} = \frac{\partial U_h/\partial D_{hj}}{\partial U_h/\partial D_{hk}} = \frac{p_j}{p_k}, \tag{4-10}$$

其中，$i = 1,\cdots,q$，$h = 1,\cdots,n$. 上式表明，在追求效用最大及完善的市场机制下可达消费品在消费者中的帕雷托最优配置.

4. 消费品帕雷托最优配置的非唯一性

一般情况下，消费品在消费者中的帕雷托最优配置点有无穷多个.

4.2 有限生产要素及消费品配置的马克思最优境界

无论资本主义市场经济还是社会主义市场经济，都应使生产要素及消费品的配置达帕雷托最优境界. 称帕雷托最优境界为有“效益”的配置. 但帕雷托最优点一般有无穷多个点，其中并非所有点都符合马克思的按劳分配的公平准则. 为此，把既符合按劳分配的公平境界，又符合帕雷托最优的有效益境界的配置称为**马克思最优境界**.

1. 按劳分配境界的定量描述

设一个人所贡献的社会必要劳动时间为 I，他所分得的产品量为 $x_1,\cdots,x_n$，每单位产品凝结的社会必要劳动时间为 $w_1,\cdots,w_n$，那么他所分得的各种产品所凝结的总时间为

$$\sum_{i=1}^{n} w_i x_i = I^*. \tag{4-11}$$

当 I 与 I^* 成正比时，便认为达到按劳分配境界.

要达到按劳分配境界，必须计算凝结在产品中的社会必要劳动时间. 对不同类型的生产函数有不同的计算方法. 由于生产函数仅是现实生产过程的近似描述，因此由它计算出的凝结在产品中的劳动价值仅是马克思劳动价值观的近似逼近.

2. 产品的劳动时间计算方法

(1) 生产过程采用静态列昂惕夫投入产出生产函数描述时，凝结在产品中劳动时间的计算方法.

静态投入产出生产函数假设每种产品仅由 1 种技术生产出来，且投入的生产要素在 1 个周期使用后立即耗尽，投人的劳动力的素质相同.

设共有 n 种产品，每生产 1 单位第 i 种产品所消耗的各种产品量为 $a_{1i}, a_{2i}, \cdots$,

a_{ni},同时消耗劳动工时为 l_i.用 w_i 表示凝结在1单位第 i 种产品中的劳动时间,那么有如下方程:

$$w_i = w_1 a_{1i} + w_2 a_{2i} + \cdots + w_n a_{ni} + l_i, \quad i = 1, \cdots, n.$$

或简记为

$$\boldsymbol{w} = \boldsymbol{w}\boldsymbol{A} + \boldsymbol{l}, \tag{4-12}$$

其中,$\boldsymbol{w} = [w_1, \cdots, w_n]$,$\boldsymbol{l} = [l_1, \cdots, l_n]$,$\boldsymbol{A}$ 为消耗系数阵.

依(4-12)式可求出凝结在各产品中的劳动时间为

$$\boldsymbol{w} = \boldsymbol{l}(\boldsymbol{I} - \boldsymbol{A})^{-1}. \tag{4-13}$$

(2) 生产过程采用具有联合产出的冯·诺伊曼生产活动分析模型描述时,凝结在产品中的劳动时间的计算方法.

设共有 n 种生产活动,第 i 种生产活动投入的各种产品量为 $a_{1i}, a_{2i}, \cdots, a_{ni}$,还投入 l_i 单位劳动工时;投入的要素经加工生产后余下的要素量为 $b_{1i}, b_{2i}, \cdots, b_{ni}$,同时产出各种产品量为 $g_{1i}, g_{2i}, \cdots, g_{ni}$.如果凝结在1单位第 i 种产品中的劳动时间为 w_i,那么有如下平衡方程:

$$\begin{aligned} & w_1 g_{1i} + w_2 g_{2i} + \cdots + w_n g_{ni} + w_1 b_{1i} + w_2 b_{2i} + \cdots + w_n b_{ni} \\ = \ & w_1 a_{1i} + w_2 a_{2i} + \cdots + w_n a_{ni} + l_i, \quad i = 1, \cdots, n. \end{aligned}$$

上式简记为

$$\boldsymbol{w}\boldsymbol{G} + \boldsymbol{w}\boldsymbol{B} = \boldsymbol{w}\boldsymbol{A} + \boldsymbol{l}, \tag{4-14}$$

其中

$$\boldsymbol{G} = \begin{bmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & & \vdots \\ g_{n1} & \cdots & g_{nn} \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}.$$

(3) 生产过程采用多技术线性多部门模型描述时,凝结在产品中劳动时间的计算方法.

设共有 n 种产品由 m 种生产活动来生产,$m \geqslant n$,即1种产品可能由多种活动来生产.投入系数阵 $\boldsymbol{A}$ 是 $n \times m$ 阵,m 列中每1列的数据为该种生产活动所投入的 n 种产品的量.产出系数阵 $\boldsymbol{G}$ 也是 $n \times m$ 阵,它的每1列的数据为该种生产活动产出的各产品的量.剩余系数阵为 $\boldsymbol{B}$,它是投入的 $\boldsymbol{A}$ 阵在1周期生产后的剩余.$\boldsymbol{d}^i = [0, \cdots, 0, 1, 0, \cdots, 0]^{\mathrm{T}}$ 为最终产出向量,其第 i 个分量为1,表示第 i 种产品产出量为1.$\boldsymbol{l}$ 为 m 维行向量,其每1分量表示该种生产活动产出量为1个单位时的劳动工时投入量.由于存在多种技术,因此1单位第 i 种产品 $\boldsymbol{d}^i$ 可以由不同的生产活动来生产.如果各种生产活动的产出为 $\boldsymbol{x}^i$,那么存在1个 $\boldsymbol{x}^i$ 使得投入劳动时间 $\boldsymbol{l}\boldsymbol{x}^i$ 最小,将 $w_i = \boldsymbol{l}\boldsymbol{x}^i$ 作为凝结在1单位第 i 种产品的劳动时间.综上所示,w_i 是如下线性规划模型的解:

$$\begin{cases} \min & w_i = \boldsymbol{l}\boldsymbol{x}^i, \\ \text{s.t.} & (\boldsymbol{G} + \boldsymbol{B} - \boldsymbol{A})\boldsymbol{x}^i \geqslant \boldsymbol{d}^i, \end{cases} \tag{4-15}$$

其中,$i = 1, \cdots, n$.

以上仅给出采用简单的线性多部门生产函数时计算凝结在产品中劳动时间的

计算方法.当投入的劳动不同质,含有简单劳动与复杂劳动时,或者采用投入要素可互相替代的生产函数时也可以有相应计算公式.但由于现实生活中的生产过程是十分复杂的,因此要十分准确地计算凝结在一种产品中的社会必要劳动时间是不可能的,只能作近似计算.

例 1 考虑只有两种要素两种消费品以及两个消费者的生产要素与消费品的帕雷托最优配置与马克思最优配置.

设有 $K_0 = 1$ 单位土地与 $L_0 = 1$ 单位劳动工时用来生产两种产品.两种产品的生产函数分别为

$$\begin{cases} x_1 = (K_1 L_1)^{1/4}; \\ x_2 = (K_2 L_2)^{1/4}. \end{cases} \tag{4-16}$$

两种要素在两个生产过程中的分配应达帕雷托最优.依(4-4)式得

$$\begin{cases} \dfrac{\partial x_1/\partial K_1}{\partial x_1/\partial L_1} = \dfrac{\partial x_2/\partial K_2}{\partial x_2/\partial L_2}, \\ K_1 + K_2 = K_0 = 1, \\ L_1 + L_2 = L_0 = 1; \end{cases} \quad 或 \quad \begin{cases} \dfrac{K_1}{L_1} = \dfrac{K_2}{L_2} = \dfrac{K_1 + K_2}{L_1 + L_2} = 1, \\ K_1 + K_2 = 1, \\ L_1 + L_2 = 1. \end{cases} \tag{4-17}$$

上式与(4-16)式共有 $x_1, x_2, K_1, K_2, L_1, L_2$ 等6个变量,消去其它变量可得到如下**生产可能性曲线方程**:

$$x_1^2 + x_2^2 = 1. \tag{4-18}$$

上式是当生产安排在帕雷托最优轨迹上两种产品产量所满足的关系式.一般情况下,生产可能性曲线难以用具体的数学表达式表示出来.

生产出的两种产品 x_1 与 x_2 在两个消费者中进行分配.设两个消费者效用函数分别为

$$\begin{cases} U_1 = (x_{11})^{1/6}(x_{12})^{1/3}, \\ U_2 = (x_{21})^{1/6}(x_{22})^{1/3}. \end{cases} \tag{4-19}$$

其中,x_{ij} 为第 i 个消费者分配到的第 j 种消费品的量.

在生产可能性曲线上任取1点(x_1^E, x_2^E),把它分配给两个消费者,消费品在消费者中配置达帕雷托最优应满足如下条件:

$$\begin{cases} \dfrac{\partial U_1/\partial x_{11}}{\partial U_1/\partial x_{12}} = \dfrac{\partial U_2/\partial x_{21}}{\partial U_2/\partial x_{22}}, \\ x_{11} + x_{21} = x_1^E, \\ x_{12} + x_{22} = x_2^E. \end{cases} \tag{4-20}$$

上式与(4-19)式一起可得到如下的**消费可能性曲线方程**:

$$U_1^2 + U_2^2 = (x_1^E)^{1/3}(x_2^E)^{2/3}. \tag{4-21}$$

生产要素及消费品的配置应使得消费者效用最大.因此应让配置处于消费可能性曲线的包络线上.在本例中,应选择(x_1^E, x_2^E)点使得消费可能性曲线半径最大,即

$$\begin{cases} \max \quad (x_1^E)^{1/3}(x_2^E)^{2/3}, \\ \text{s.t.} \quad (x_1^E)^2 + (x_2^E)^2 = 1. \end{cases} \tag{4-22}$$

由上式求出

$$(x_1^E)^* = (1/3)^{1/2}, \quad (x_2^E)^* = (2/3)^{1/2}.$$

将最优解$(x_1^E)^*$,$(x_2^E)^*$代入(4-21)式,得

$$U_1^2 + U_2^2 = 0.7274159. \tag{4-23}$$

上式是消费可能性曲线的包络线.它既符合生产要素的帕雷托最优配置,又符合消费品的帕雷托最优配置.称(4-23)式为生产要素及消费品有"效益"的配置.但在(4-23)式中仍有无穷多个点,而其中只有1点符合马克思的按劳分配公平准则,即其中只有1点为马克思最优配置点.

当生产安排在帕雷托最优配置时,两种产品的产量分别为$(x_1^E)^*$与$(x_2^E)^*$,依生产函数(4-16)式及(4-17)式,有

$$\begin{cases}(x_1^E)^* = (K_1 L_1)^{1/4} = (L_1 \times L_1)^{1/4} = (L_1)^{1/2} = (1/3)^{1/2},\\ (x_2^E)^* = (K_2 L_2) = (L_2 \times L_2)^{1/4} = (L_2)^{1/2} = (2/3)^{1/2}.\end{cases}$$

从上式可求出凝结在各种产品中的劳动时间为

$$V_1 = L_1/(x_1^E)^* = (1/3)^{1/2},$$

$$V_2 = L_2/(x_2^E)^* = (2/3)^{1/2}.$$

设第1个消费者付出的劳动时间为A,第2个消费者付出的劳动时间为B,当他们分得的消费品中所凝结时间等于他们各自付出的时间时,达按劳分配境界.这时应满足如下方程:

$$\begin{cases}V_1 x_{11} + V_2 x_{12} = A,\\ V_1 x_{21} + V_2 x_{22} = B.\end{cases} \tag{4-24}$$

一旦给出A与B的具体数值,便可找到既公平又有效益的马克思最优配置点.比如,若第1个消费者同时又是第1种产品的劳动者,第2个消费者同时又是第2种产品的生产者,即$A = L_1 = 1/3, B = L_2 = 2/3$,依以上各式可求出马克思最优配置点为

$$\begin{cases}x_{11} = 1/(3\sqrt{3}), \quad x_{12} = \sqrt{2}/(3\sqrt{3}),\\ x_{21} = 2/(3\sqrt{3}), \quad x_{22} = 2\sqrt{2}/(3\sqrt{3}).\end{cases} \tag{4-25}$$

以上通过计算的方法或"计划"的方法求出了马克思最优配置点.4.1节曾指出市场调节可达消费品与生产要素的帕雷托最优境界.那么,一个十分重要的问题是:市场调节能达到马克思最优配置点吗?

现假设通过市场调节来实现生产要素与消费品的配置.在市场机制下,本例中存在如下几种人:地主(土地所有者)、生产经营者(不付出劳动但占有利润)、生产过程中付出劳动者即工人或农民(占有工资)、消费者.在更复杂的模型中还应包括资本家,即资本的占有者.设土地租金为r、劳动工时的工资率为w、产品市场价格为p_1与p_2.

生产经营者在利润最大法则之下的产品供给函数为

$$x_1 = p_1/(4\sqrt{rw}), \tag{4-26}$$

$$x_2 = p_2/(4\sqrt{rw}), \tag{4-27}$$

其中,产品供给量 x_1, x_2 为市场价格的函数;投入要素需求函数为

$$K_1 = p_1^2/(16r\sqrt{rw}), \tag{4-28}$$

$$K_2 = p_2^2/(16r\sqrt{rw}), \tag{4-29}$$

$$L_1 = p_1^2/(16w\sqrt{rw}), \tag{4-30}$$

$$L_2 = p_2^2/(16w\sqrt{rw}), \tag{4-31}$$

其中,要素投入量 K_1, K_2, L_1, L_2 均为市场价格的函数;获得的利润函数为

$$\Pi = (p_1^2 + p_2^2)/(8\sqrt{rw}), \tag{4-32}$$

其中,Π 为两种产品生产过程中所获得的总利润.

设消费者效用函数都一样,如(4-19)式所示,在效用最大法则之下可得到相应的需求函数.

第1种产品的生产者付出劳动 L_1,获得工资 wL_1,作为消费者所需求的两种产品的量 x_{11} 与 x_{12} 由如下需求函数描述:

$$p_1x_{11} = wL_1/3, \tag{4-33}$$

$$p_2x_{12} = 2wL_2/3. \tag{4-34}$$

第2种产品的生产者付出劳动 L_2,获得工资 wL_2,作为消费者所需求的两种产品的量 x_{21} 与 x_{22} 由如下需求函数描述:

$$p_1x_{21} = wL_2/3, \tag{4-35}$$

$$p_2x_{22} = 2wL_2/3. \tag{4-36}$$

地主占有1单位土地,获利为 r,作为消费者对两种产品的需求量 x_{31} 与 x_{32} 由如下需求函数描述:

$$p_1x_{31} = r/3, \tag{4-37}$$

$$p_2x_{32} = 2r/3. \tag{4-38}$$

生产经营者获得利润 Π,作为消费者对两种产品的需求量 x_{41} 与 x_{42} 由如下需求函数描述:

$$p_1x_{41} = \frac{1}{3}\Pi, \tag{4-39}$$

$$p_2x_{42} = \frac{2}{3}\Pi. \tag{4-40}$$

要素市场供求平衡由下式描述:

$$K_1 + K_2 = 1, \tag{4-41}$$

$$L_1 + L_2 = 1. \tag{4-42}$$

产品市场供求平衡由下式描述:

$$x_{11} + x_{21} + x_{31} + x_{41} = x_1, \tag{4-43}$$

$$x_{12} + x_{22} + x_{32} + x_{42} = x_2. \tag{4-44}$$

自(4-26)式 ~ (4-44)式,共有19个方程、19个未知数,但只有18个方程相互独立.因此可设工资率 $w = 1$,那么上述市场供求系统解为

$$r = w = 1,$$
$$p_1 = 4/\sqrt{3}, p_2 = 4\sqrt{6}/3,$$
$$x_1 = (1/3)^{1/2}, x_2 = (2/3)^{1/2},$$
$$K_1 = 1/3, L_1 = 1/3,$$
$$K_2 = 2/3, L_2 = 2/3.$$

从以上分析可知，通过市场调节使生产的安排恰好处在帕雷托最优境界上，供给量 $x_1 = (x_1^E)^*$，$x_2 = (x_2^E)^*$. 全社会生产产品中所凝结的劳动时间等于劳动者贡献的时间，即

$$V_1 x_1 + V_2 x_2 = L_1 + L_2.$$

从(4-43)式、(4-44)式可知，劳动者分配到的消费品小于其贡献的时间，即

$$V_1(x_{11} + x_{21}) + V_2(x_{12} + x_{22}) < (V_1 x_1 + V_2 x_2) = L_1 + L_2,$$

这意味着完善的市场经济达不到生产要素与消费品的马克思最优配置.

但是，如果劳动者同时又是土地的拥有者及生产经营者，而且土地拥有量及利润分配与其贡献的劳动时间成比例，那么市场调节可达马克思最优境界. 不过这是一种极特殊的情况，在目前的现实生活中几乎不可能出现.

本例的分析可推广到更一般的情况，并得到如下重要结论：存在地主（资源私有者）、资本家（资本占有者）、生产经营者（如包工头等）的情况下，完善的市场经济可达生产要素及消费品的帕雷托最优境界，但达不到马克思最优境界.

4.3 可持续最优发展的目标设定

设 $U(t)$ 表示第 t 年某个国家或地区的目标值，它应是该国家或地区各种人各自目标值的函数，即

$$U(t) = U_0[U_{1(t)}(t), U_{2(t)}(t), \cdots, U_{j(t)}(t), \cdots, U_{m(t)}(t)], \tag{4-45}$$

其中，$j(t)$ 表示第 t 时期的第 j 种人，$j = 1, \cdots, m$.

(4-45)式中 $U_{j(t)}(t)$ 不应是第 t 时期第 j 种人自己所感觉到的实际效用函数. 因为每个人效用函数值的大小是难以度量的，因此它应是目标设定者所给出的对第 j 种人所获得的效用值的价值判断准则.

经济系统从 t 时期开始直至以后无穷的可持续发展总目标 $U_\Sigma(t)$ 是各时期目标值的函数，即

$$U_\Sigma(t) = U_\Sigma[U_0(t), U_1(t + \Delta t), \cdots, U_i(t + i\Delta t), \cdots]. \tag{4-46}$$

(4-46)式是人们利用 t 时期之前的已有知识去推测未来世界发展并构造出的目标函数或价值判断准则. 随着时间的推移，人们知识与观念发生变化，价值判断准则也将作相应变化.

具体给出(4-46)式的数学表达式属于规范经济学的范畴，如4.1节与4.2节所述，$U_\Sigma(t)$ 的设定应使得经济系统到达马克思最优境界. 而且，不仅应使得一代人之间达到按劳分配的公平境界，还应考虑这一代人与下一代人之间的公平问题，即应考虑所谓的代间与代际公平问题.

例2 设第 t 时期目标值 $U(t)$ 只与人均消费 $C(t)$ 有关. 总目标值 $U_{\Sigma}(t)$ 是当前消费与未来消费之间的加权和,即

$$U_{\Sigma}(t) = \lambda(t)\ C(t)\ \Delta t + \lambda(t+\Delta t)\ C(t+\Delta t)\Delta t + \cdots + \lambda(t+i\Delta t)\ C(t+i\Delta t)\Delta t + \cdots,$$

其中,$\lambda(t+i\Delta t)$ 为权重. 当 $\Delta t \to 0$ 时,有

$$U_{\Sigma}(t) = \int_{t}^{\infty}\lambda(t)\ C(t)\ \mathrm{d}t, \tag{4-47}$$

上式中 $\lambda(t)$ 通常取 e^{-rt} 的形式,即

$$U_{\Sigma}(t) = \int_{t}^{\infty}C(t)\ \mathrm{e}^{-rt}\ \mathrm{d}t, \tag{4-48}$$

其中,r 可理解为未来消费在 t 时刻的贴现率.

需要指出,(4-48)式所示目标函数值没有考虑代间与代际公平问题,也没有考虑环境、国防安全、健康等更广泛、深入的"效益"问题.

5 可持续最优经济发展轨道

5.1 线性多部门经济系统的最优增长与快车道定理

考虑如下的冯·诺伊曼-列昂惕夫模型:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t+1) + \boldsymbol{T}\boldsymbol{x}(t+1), \tag{5-1}$$

其中,$\boldsymbol{x}(t)$ 为 n 维产出向量,$\boldsymbol{A}$ 为消耗系数阵,$\boldsymbol{T}$ 为消费系数阵.

(5-1)式所示系统在 $t=0$ 时各部门产出向量或产出结构为已知值 $\boldsymbol{x}(0)$. 在 τ 个时间周期之后的实际产出结构为 $\boldsymbol{x}(\tau)$,而希望的产出结构为 $\boldsymbol{y} = [y_1,\cdots,y_n]^{\mathrm{T}}$,其中 $y_1+\cdots+y_n = 1$.

定义

$$q_i = x_i(\tau)/y_i,\quad i = 1,\cdots,n, \tag{5-2}$$

其中,$x_i(\tau)$ 是 $\boldsymbol{x}(\tau)$ 的第 i 个分量.

系统目标是在给定的产出终端结构下各部门的产出尽可能大,即目标 J 为

$$\max J,\quad q = \min\{q_1,\cdots,q_n\}. \tag{5-3}$$

在(5-1)式中,记 $\boldsymbol{H} = \boldsymbol{A} + \boldsymbol{T}$ 为消耗、消费系数阵. 为在第 $t+1$ 时间周期产出 $\boldsymbol{x}(t+1)$,需在第 t 个周期投入消耗及消费品的量为 $\boldsymbol{H}\boldsymbol{x}(t+1)$. 由于第 t 周期产出为 $\boldsymbol{x}(t)$,因此应满足如下约束:

$$\boldsymbol{H}\boldsymbol{x}(t+1) \leqslant \boldsymbol{x}(t). \tag{5-4}$$

依(5-2)式定义,在终端时期应满足如下不等式约束:

$$q\,\boldsymbol{y} \leqslant \boldsymbol{x}(\tau). \tag{5-5}$$

在(5-4)式、(5-5)式约束下,(5-3)式目标 J 取极大的数学模型为如下的线性规划问题:

$$\begin{cases}\max & q,\\ \text{s.t.} & \boldsymbol{H}\boldsymbol{x}(1)\leqslant \boldsymbol{x}(0),\\ & \boldsymbol{H}\boldsymbol{x}(2)\leqslant \boldsymbol{x}(1),\\ & \vdots\\ & \boldsymbol{H}\boldsymbol{x}(\tau)\leqslant \boldsymbol{x}(\tau-1),\\ & q\boldsymbol{y}\leqslant \boldsymbol{x}(\tau),\\ & \boldsymbol{x}(1),\cdots,\boldsymbol{x}(\tau)\geqslant 0, q\geqslant 0.\end{cases}\tag{5-6}$$

把上式写成矩阵形式,有

$$\begin{cases}\max\quad q,\\ \text{s.t.}\begin{bmatrix}\boldsymbol{H} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ -\boldsymbol{I} & \boldsymbol{H} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \vdots & \vdots & & \vdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{H} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & -\boldsymbol{I} & \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(1)\\ \boldsymbol{x}(2)\\ \vdots\\ \boldsymbol{x}(\tau)\\ q\end{bmatrix}\leqslant\begin{bmatrix}\boldsymbol{x}(0)\\ \boldsymbol{0}\\ \vdots\\ \boldsymbol{0}\end{bmatrix},\\ \quad[\boldsymbol{x}^{\mathrm{T}}(1),\cdots,\boldsymbol{x}^{\mathrm{T}}(\tau),q]^{\mathrm{T}}\geqslant 0.\end{cases}\tag{5-7}$$

在上式中,记

$$\boldsymbol{x}=[\boldsymbol{x}^{\mathrm{T}}(1),\cdots,\boldsymbol{x}^{\mathrm{T}}(\tau),q]^{\mathrm{T}},$$

$$q=[\boldsymbol{0},\cdots,\boldsymbol{0},1]\begin{bmatrix}\boldsymbol{x}(1)\\ \boldsymbol{x}(2)\\ \vdots\\ \boldsymbol{x}(\tau)\\ q\end{bmatrix}=\boldsymbol{c}\boldsymbol{x},$$

$$\widetilde{\boldsymbol{H}}=\begin{bmatrix}\boldsymbol{H} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ -\boldsymbol{I} & \boldsymbol{H} & \cdots & \boldsymbol{0} & \boldsymbol{0}\\ \vdots & \vdots & & \vdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{H} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & -\boldsymbol{I} & \boldsymbol{y}\end{bmatrix},$$

$$\boldsymbol{b}=[\boldsymbol{x}^{\mathrm{T}}(0),\boldsymbol{0},\cdots,\boldsymbol{0}]^{\mathrm{T}}.$$

则(5-7)式化为标准的线性规划问题.

$$\begin{cases}\max & \boldsymbol{c}\boldsymbol{x},\\ \text{s.t.} & \widetilde{\boldsymbol{H}}\boldsymbol{x}\leqslant \boldsymbol{b},\\ & \boldsymbol{x}\geqslant 0.\end{cases}\tag{5-8}$$

对(5-7)式或(5-8)式的线性规划问题有如下的快车道定理:如果 $\boldsymbol{H}$ 是不可分解非负本原方阵,且经济系统初始结构 $\boldsymbol{x}(0)$ 及终端目标结构 $\boldsymbol{y}$ 都是严格正的,当规划时间周期 τ 足够大时,在 τ 的大部分时间内,经济系统产出比例或产出结构都充分接近 $\boldsymbol{x}^*$,而 $\boldsymbol{x}^*$ 为(5-1)式所示冯·诺伊曼-列昂锡夫模型平衡增长轨道上的比例,即 $\boldsymbol{x}^*$ 符合下式:

$$\boldsymbol{x}^*=(1+\lambda)\boldsymbol{H}\boldsymbol{x}^*,\tag{5-9}$$

式中,λ 为平衡增长率.$\boldsymbol{x}^*$ 称为冯·诺伊曼射线.

快车道定理表明,当产出比例 $x(0)$ 不在冯·诺伊曼射线上时,应尽快调整到冯·诺伊曼射线上,然后沿此射线快速前进,最后再尽快调整到终端比例 y 上.通常将上述冯·诺伊曼射线称为"快车道".

"快车道"定理最早由陶夫曼(R. Dofman)、萨缪尔森(P. A. Samuelson)、索洛(Solow)等人利用线性规划应用于线性多部门模型,其后日本的森岛等人对"快车道"定理深入研究,并使之适用于各种类型的线性多部门模型.例如,若将(5-1)式的冯·诺伊曼-列昂惕夫模型改为

$$\boldsymbol{B}x(t) = \boldsymbol{A}x(t+1) + \boldsymbol{T}x(t+1), \tag{5-10}$$

那么只须将(5-7)式中的单位阵 $\boldsymbol{I}$ 改为 $\boldsymbol{B}$ 阵便可得到相应的经济系统长期规划模型及相应结论.

在上述经济增长"快车道"模型中,由于投入的劳动力素质相同,因此没有考虑按劳分配的"公平"问题.此外,当规划的时间 τ 很长时,也没有考虑当前这一代人消费与下一代人消费的代际公平问题.对上述问题的进一步深入研究,要构造更加复杂的线性多部门模型并采用更加切合实际的目标函数.例如,可以采用多种技术有联合产出且投入劳动力素质不同,并考虑人才培养的线性多部门模型,通过计算凝结在产品及不同素质人才中的劳动时间来分析如何达按劳分配的公平境界.

5.2 动态经济系统最优经济策略设计及最优发展轨道的计算

中国是一个发展中国家,从当前状态过渡到基本实现现代化的目标状态有几个基本问题:

(1) 存在性　即从当前状态到目标状态是否存在一条最优发展轨道.

(2) 可计算性　即能否对最优轨道给出有一定准确度的计算方法.

(3) 求解的必要性　各种宏观调控策略的正确使用就是要使中国经济沿最优轨道前进.如果对最优轨道不能给出有一定准确度的定量描述,那就意味着无法施加正确的经济调控策略.

由于经济系统的复杂性,要十分准确计算经济系统的最优轨道是不可能的,但对最优轨道给出有一定意义的较为准确的计算结果不仅可能,也是必要的.

求解最优经济增长轨道主要包括以下两方面的工作:

(1) 经济系统目标的定量化描述　即对效益与公平给出定量化描述."效益"指标包括人均GDP长期可持续增长速度、生态环境、国防安全、身体健康等."公平"指标包括一代人之间、各代人之间的公平准则及按劳分配境界的定量描述.总之,要给出既有效益又有公平的马克思最优境界的定量描述.

(2) 构造从策略变量到目标变量间因果关系链的数学模型　经济系统策略变量的确定与经济体制有关.在市场经济体制下,主要有货币政策与财政政策两大类策略变量.货币政策主要有货币发行量等变量.财政政策包括财政收入与财政支出两方面.在财政收入方面主要有各种税的税率,如增值税率、企业所得税率、个人所得税率、关税率等.财政支出则主要有用于国防支出、环境保护及公共投资等方面的支出,财政支出政策主要决定各项支出的最优比例.

构造市场经济的模型应考虑货币市场、劳动市场、资本市场、产品市场、资源市场等几个基本市场，各个市场的供给函数与需求函数是市场价格及各政策变量的函数．例如，在产品市场中产品供给函数是产品价格及增值税率、企业所得税率等的函数，产品需求函数是产品市场价格及工资收入、利息率、个人所得税率等的函数．各个市场的供求函数相互关联构成复杂的大规模经济系统．系统涉及的主要变量有产品产量、需求量、进出口量、产品价格，增值税、企业所得税、关税等各种税的税率，利率，汇率，人口，就业人口，国民生产总值(GNP)，国内生产总值(GDP)，等等．

一旦给出经济系统的数学模型及相应的目标函数，便可以依据有关的数学工具求解经济发展的最优轨道．例如，可以利用庞得里亚金极大值原理求解动态经济系统的最优发展轨道．

应当指出，由于经济系统的复杂性，人们难以用完善的数学模型来准确地描述现实世界的运行规律．例如，在生产过程中除了产出产品还排放污染，污染物对环境产生破坏作用，环境的变化又反过来影响了生产的发展．而要用数学方程来描述环境变化对人类可持续发展的影响，则需要长时间的数据积累．由于在现实的经济中人们缺乏实际数据的积累，因而现在无法用准确的数学公式来描述环境变化对生产的影响．但是，不可能建立现实经济系统十分准确的数学模型并不意味着不可能建立相应较为准确的数学模型．随着经济控制论理论与实践的深入发展，人们必将采用各种更加有效的调控策略，让现实经济系统沿着最优发展轨道前进，直奔美好的目标境界．

例1 地区经济增长快车道模型．

假设在其它条件不变的情况下，某地区工业生产净值 Y 与交通运输能力 K_1（由交通部门的资本存量衡量）及生产部门资本存量 K_2 有关，可由如下式子表示：

$$Y = A K_1^a K_2^b. \tag{5-11}$$

净产出 Y 中有 σY 用于运输部门及生产部门的投资．而 σY 中有 $\alpha(t)\sigma Y$ 用于交通运输部门的投资．设交通部门资本存量 K_1 的增加由下式描述：

$$dK_1/dt = \alpha(t)\sigma Y. \tag{5-12}$$

σY 中有 $(1-\alpha(t))\sigma Y$ 用于生产部门资本存量的投资，K_2 的增加由下式描述：

$$dK_2/dt = [1-\alpha(t)]\sigma Y. \tag{5-13}$$

开始时，各部门资本存量为 $K_1(0)$ 与 $K_2(0)$．系统的目标是使得累积净产出最大．如果规划时间为 T，那么本例数学模型为

$$\begin{cases} \max\limits_{0\leqslant\alpha(t)\leqslant1} J = \int_0^T Y(t)dt; \\ \text{s.t.}\quad dK_1/dt = \alpha(t)\sigma Y(t), \\ \qquad dK_2/dt = [1-\alpha(t)]\sigma Y(t), \\ \qquad K_1(0), K_2(0) \text{ 给定}. \end{cases} \tag{5-14}$$

可用庞得里亚金极大值原理求解(5-14)式的最优轨道．

作哈密尔顿函数：

$$H = Y(t) + \lambda_1(t)\ \alpha(t)\ \sigma Y(t) + \lambda_2(t)\ [1 - \alpha(t)]\sigma Y(t).$$

(5-14) 式的极值必要条件为

$$\begin{cases} \mathrm{d}\lambda_1/\mathrm{d}t = -\partial H/\partial K_1; \\ \mathrm{d}\lambda_2/\mathrm{d}t = -\partial H/\partial K_2; \\ H(K_1^*, K_2^*, \alpha^*, \lambda_1^*, \lambda_2^*) = \max\limits_{0\le\alpha(t)\le1} H(K_1^*, K_2^*, \alpha, \lambda_1^*, \lambda_2^*); \\ \mathrm{d}K_1/\mathrm{d}t = \alpha(t)\sigma Y(t); \\ \mathrm{d}K_2/\mathrm{d}t = [1 - \alpha(t)]\sigma Y(t); \\ K_1(0), K_2(0)\ \text{给定}. \end{cases} \tag{5-15}$$

上式第三个必要条件表明,应取 $\alpha(t)$ 使哈密尔顿函数值 H 最大. 由于 $Y(t) > 0$,故应使 $1 + \lambda_1(t)\ \alpha(t)\ \sigma + \lambda_2(t)[1 - \alpha(t)]\sigma$ 取极大. 由于

$$1 + \lambda_1(t)\alpha(t)\sigma + \lambda_2(t)[1 - \alpha(t)]\sigma = 1 + [\lambda_1(t) - \lambda_2(t)]\alpha(t)\sigma + \lambda_2(t)\sigma,$$

因此最优策略为

$$\begin{cases} \text{当}\ \lambda_1(t) > \lambda_2(t)\ \text{时}, & \alpha(t) = 1; \\ \text{当}\ \lambda_1(t) = \lambda_2(t)\ \text{时}, & \alpha(t)\ \text{待定}; \\ \text{当}\ \lambda_1(t) < \lambda_2(t)\ \text{时}, & \alpha(t) = 0. \end{cases} \tag{5-16}$$

当 $\lambda_1(t) = \lambda_2(t)$ 时,由(5-15) 式的第一、二个必要条件得

$$\partial Y/\partial K_1 = \partial Y/\partial K_2,$$

将(5-11) 式代入上式,得

$$K_1 : K_2 = a : b,$$

或

$$\mathrm{d}K_1/\mathrm{d}t : \mathrm{d}K_2/\mathrm{d}t = a : b. \tag{5-17}$$

由式(5-12)、(5-13)、(5-17) 得

$$\alpha(t) = a/(a + b).$$

由于最优策略只能取 $\alpha(t) = 1$,$\alpha(t) = 0$ 及 $\alpha(t) = a/(a + b)$,因此最优策略如下:

情况 1　当交通部门资本存量 K_1 规模太小,或交通负荷过重时,即 $K_1/a < K_2/b$ 时,可将全部投资用于交通建设,直至达到合适比例为止,这时策略为 $\alpha(t) = 1$.

情况 2　当交通部门资本存量 K_1 规模太大,这时可暂不对交通部门投资,即 $\alpha(t) = 0$, 全部的 σY 投资于生产部门,直至达合适比例.

情况 3　当交通部门及生产部门资本存量成合适比例,即 $K_1/a = K_2/b$ 时,交通部门投资比例为 $\alpha(t) = a/(a + b)$,生产部门投资为 $1 - \alpha(t) = b/(a + b)$.

称情况 3 为经济增长的"快车道". 当交通部门及生产部门资本存量比例成恰当值时,经济系统处于良性增长轨道;当不在这个比例上时,应尽快调整到恰当比例上.

例 2　考虑如下宏观经济模型:

总需求方程

$$D = C + I + G. \tag{5-18}$$

其中，D 为总需求；I 为投资总需求，它包括公共投资 I_1 及企业与个人投资 I_2；G 为政府购买，主要用于公共消费.

生产函数

$$Y = AK_1^a K_2^b L^{1-a-b} e^{\beta t}. \tag{5-19}$$

其中，Y 为实际国民生产总值；A 为常数；β 为技术进步系数；a，b 为参数；K_1 为公共固定资本存量；K_2 为企业及个人固定资本存量；L 为就业人口.

名义国民生产总值(GNP)是实际国民生产总值 Y 与市场物价指数 $(1+\tau)p$ 相乘的结果，即

$$\text{GNP} = (1+\tau)pY, \tag{5-20}$$

其中，τ 为增值税率，p 为应税价格.

$$\text{应税增加值} = \text{GNP} \div (1+\tau) = pY. \tag{5-21}$$

应税增加值是增值税的税基. 即增值税收入为应税增加值乘以税率 τ. 用公式表示即为

$$\text{增值税收入} = \tau pY. \tag{5-22}$$

本模型只考虑 1 个税种(可推广到多税种情况)，因此可以认为税收收入总额即为增值税收入额，即

$$\text{税收收入总额 } \tilde{T} = \text{增值税收入额} = \tau pY. \tag{5-23}$$

如下的可支配收入可用于新增资本投资及抵消固定资本折旧的重置投资，余下的用于个人消费：

$$\text{可支配收入} = \text{GNP} - \text{税收收入总额} = pY. \tag{5-24}$$

在市场机制下，企业及个人投资总额 I_2(包括新增固定资本投资及抵消固定资本折旧的重置投资)是可支配收入 pY 及利率 r 的函数，设由如下方程描述：

$$I_2 = I_2(r)pY, \tag{5-25}$$

其中，$I_2(r)$ 是 r 的递减函数.

个人消费 C 也可看做是可支配收入 pY 及利率 r 的函数，设由如下方程描述：

$$C = C(r)pY. \tag{5-26}$$

设财政支出总额为 T，其中一部分 $T \times g$ 用于公共消费 G，另一部分 $T(1-g)$ 用于公共投资 I_1，即

$$G = gT, \tag{5-27}$$

$$I_1 = (1-g)T, \tag{5-28}$$

上两式中，T，g 都是政策变量或外生变量.

对于公共固定资本 K_1 来讲，它的增加等于名义投资 I_1 除以市场价格 $(1+\tau)p$ 得到的实际公共投资 $I_1/(1+\tau)p$ 减去公共固定资本实际折旧额 $\delta_1 K_1$，即

$$\frac{dK_1}{dt} = \frac{I_1}{(1+\tau)p} - \delta_1 K_1, \tag{5-29}$$

其中，δ_1 为折旧率.

对于企业及个人拥有的固定资本存量 K_2 来讲，它的增加为实际投资额 $I_2/(1+\tau)p$ 减去实际折旧额 $\delta_2 K_2$，即

$$\frac{dK_2}{dt} = \frac{I_2}{(1+\tau)p} - \delta_2 K_2, \tag{5-30}$$

其中,δ_2 为折旧率.

假设就业人口按一定的增长率增长,即

$$L = L_0 e^{nt}, \tag{5-31}$$

其中,n 为就业人口增长率.

市场物价由实际总需求与实际总供给之差决定,当总需求大于总供给时,物价上升;反之,物价下降.物价变化可由下述方程描述:

$$\frac{dp}{dt} = k_1\left(\frac{D}{(1+\tau)p} - Y\right), \tag{5-32}$$

其中,$k_1 > 0$ 为调节系数.

货币总需求 M^d 是利率 r 及名义国民生产总值的函数.它由如下方程描述:

$$M^d = M^d(r) \times (1+\tau)pY, \tag{5-33}$$

其中,$M^d(r)$ 是 r 的递减函数.

利率 r 的变化由货币供求差额决定:

$$\frac{dr}{dt} = k_2 \times (M^d - M^s), \tag{5-34}$$

其中,M^s 为货币供应量,它是外生政策变量.$k_2 > 0$ 为调节系数.

(5-18) ~ (5-34) 式构成一个简易的单个生产部门的宏观总量模型.

首先讨论上述模型的平衡增长轨道.

在平衡增长轨道上,公共固定资本存量 K_1 以及企业及个人固定资本存量 K_2,对产出 Y 的边际贡献应一致,即

$$\partial Y/\partial K_1 = \partial Y/\partial K_2.$$

将生产函数代入上式,得

$$K_1(t) : K_2(t) = a : b,$$

这意味着 $K_1(t)$ 与 $K_2(t)$ 有相同的增长率 x,即

$$\begin{cases} K_1(t) = K_{10}e^{xt}, \\ K_2(t) = K_{20}e^{xt}. \end{cases} \tag{5-35}$$

产出 Y 在平衡增长轨道上也有相同增长率,即

$$Y = Y_0 e^{xt}. \tag{5-36}$$

上两式中,K_{10}, K_{20}, Y_0 为常数.将(5-35)、(5-36) 式代入生产函数,求出平衡增长率 x 为

$$x = n + \beta/(1-a-b).$$

由上式可以看出,平衡增长速度一部分由就业人口扩张所引起(大小为 n),称之为外延式扩大再生产;另一部分由技术进步系数 β 所引起,称之为内涵式扩大再生产.

在平衡增长时,税收收入 $\tilde{T}$ 应等于税收支出 T,即 $\tilde{T} = T$.再利用模型方程可求出

$$\frac{Y_0}{L_0}=(1-g)^{a/(1-a-b)}\left(\frac{\tau}{1+\tau}\right)^{a/(1-a-b)}\left(\frac{1}{1+\tau}\right)^{b/(1-a-b)}\times\varepsilon, \qquad (5\text{-}37)$$

其中

$$\varepsilon=A^{1/(1-a-b)}\times\frac{[I_2(r)]^{b/(1-a-b)}}{(x+\delta_1)^{a/(1-a-b)}(x+\delta_2)^{b/(1-a-b)}}.$$

在平衡增长时,货币发行等于货币需求,故 r 为常数.因此 ε 为常数.

下面给出平衡增长轨道上的目标函数 U.它是人均公共实际消费与人均个人消费的函数,设它由如下效用函数表达:

$$U=\left\{\frac{G/(1+\tau)p}{L}\right\}^{s}\left\{\frac{C/(1+\tau)p}{L}\right\}^{w}. \qquad (5\text{-}38)$$

其中,s 与 w 为偏好系数,可设 $s+w=1$.

在平衡增长时 $G=gT=g\tilde{T}=g\tau pY_0e^{xt}$, $C=C(r)pY_0e^{xt}$,代入(5-38)式求出:

$$U=g^{s}\left(\frac{\tau}{1+\tau}\right)^{s}\left(\frac{1}{1+\tau}\right)^{w}(1-g)^{a/(1-a-b)}\left(\frac{\tau}{1+\tau}\right)^{a/(1-a-b)}\times\left(\frac{1}{1+\tau}\right)^{b/(1-a-b)}\times\varepsilon\times e^{(x-n)t}\times[C(r)]^{w}. \qquad (5\text{-}39)$$

注意到上式中 x,n,r 都不是政策变量,而 g,τ 为政策变量.要取 g,τ 使目标值 U 最大的必要条件为

$$\begin{cases}\partial U/\partial g=0,\\ \partial U/\partial\tau=0.\end{cases}$$

依上式及(5-39)式求出平衡增长轨道上最优策略为

$$\begin{cases}g=\dfrac{s\times(1-a-b)}{a+s(1-a-b)},\\[2mm] \tau=\dfrac{a+s(1-a-b)}{b+w(1-a-b)}.\end{cases}$$

上式称为经济系统运行在平衡增长轨道上的最优财政政策,即应采用什么样的最优税率 τ,以及财政收入中用于公共消费的最优比例 g.

6　经济控制论研究的进展特点

经济控制论正在不断的发展之中,它包含十分庞大的内容与众多的研究方向.各个方向上都可列出许多著名的科学家与相应的重要成果.因此要想说清经济控制论在各方向上的进展是一件极为困难的事情.但经济控制论研究的进展特点可以用“交叉性”三个字来概括,即经济学、控制理论、数学等学科之间相互交叉.下面以控制理论中代数系统理论在经济学中的应用为例来说明这种交叉性.代数系统理论是控制理论的一个研究方向,它研究域、环、格等抽象代数系统的求解与应用问题.

中国古代的公鸡、母鸡、小鸡问题:“鸡翁一值钱五,鸡母一值钱三,鸡雏三值钱

一,百钱买百鸡,问鸡翁、鸡母、鸡雏各几何?"

用 x,y,z 分别表鸡翁、鸡母、鸡雏的个数,得

$$\begin{cases} 5x + 3y + z/3 = 100, \\ x + y + z = 100. \end{cases}$$

消去 z,得

$$7x + 4y = 100. \tag{6-1}$$

上式为整数环上的不定方程或丢番图方程.整数环上二元一次不定方程可记为

$$ax + by = d, \tag{6-2}$$

其中,$a,b,d \in$ 整数环.对(6-1)式或(6-2)式的求解可在有关数论的教材中找到.

再看萨缪尔森等人的简易宏观经济模型:用 $c(t)$,$Y(t)$ 分别表示第 t 年总消费及国民收入,k_1 表示边际消费倾向,则有消费函数

$$c(t+1) = k_1 Y(t), \quad 0 < k_1 < 1. \tag{6-3}$$

用 $I(t)$,$I_1(t)$,$I_2(t)$ 分别表示第 t 年总投资、诱发性投资、自发性投资,有定义方程

$$I(t) = I_1(t) + I_2(t). \tag{6-4}$$

诱发性投资的变化与消费额的增加有关系,其关系可用如下方程描述:

$$I_1(t+1) = k_2[c(t+1) - c(t)], k_2 > 0. \tag{6-5}$$

第 t 年的国民收入 $Y(t)$ 用于投资与消费,有如下平衡方程:

$$Y(t) = I_1(t) + I_2(t) + c(t). \tag{6-6}$$

自发性投资 $I_2(t)$ 是由决策者掌握的输入变量.如果 $I_2(t)$ 增长快速,则消费额 $c(t)$ 也增长快速.由于 $c(t)$ 是名义值,$c(t)$ 增长太快意味着通货膨胀率高(本模型无价格变量).因此名义消费 $c(t)$ 不能增长太快.设 $w(t)$ 为决策者所希望的第 t 年消费额,它的增长速度应与实际产出增长速度相协调,希望的消费与现实的名义消费之差为 $e(t)$:

$$e(t) = w(t) - c(t). \tag{6-7}$$

(6-3) ~ (6-7) 式构成线性离散时间动态宏观经济系统.政策设计者的任务是确定应采取什么样的输入 $I_2(t)$ 使 $e(t)$ 尽快为零.

萨缪尔森等人给出的简易模型与中国古代的鸡翁、鸡母、鸡雏问题有什么联系呢?

考虑离散时间函数 $x(t)$,它的 Z 变换为

$$x(z) = \sum_{t=0}^{\infty} x(t) z^{-t}. \tag{6-8}$$

$x(t)$ 为时间 t 的实数域上函数,而 $x(z)$ 为移位算子环上元素.依 Z 变换知识,如果 $c(t)$ 的 Z 变换为 $c(z)$,那么 $c(t)$ 的 Z 变换为 $zc(z) - zc(0)$.若 $Y(t)$ 的 Z 变换为 $Y(z)$,则 $Y(t+1)$ 的 Z 变换为 $zY(z) - zY(0)$.再设 $I_1(t)$,$I_2(t+1)$,$I(t)$,$e(t)$ 的 Z 变换分别为 $I_1(z)$,$I_2(z)$,$I(z)$,$e(z)$,希望的消费函数 $w(t)$ 按 10% 增长,即 $w(t) = 1.1^t$,那么 $w(z) = z/(z-1.1)$.系统初始条件为 $Y(0) = 1$,$c(0) = 1$,参数 $k_1 = 0.5$,$k_2 = 0.6$,那么(6-3)式 ~ (6-7)式离散时间系统经 Z 变换后可化为如下移位算子环上的代数方程:

$$(1-0.8z^{-1}+0.3z^{-2})e(z)+0.5z^{-2}I_2(z) = \frac{1-0.8z^{-1}+0.3z^{-2}}{1-1.1z^{-1}}-(1-0.3z^{-1}). \tag{6-9}$$

令

$$\begin{cases} a(z) = 1-0.8z^{-1}+0.3z^{-2}, \\ b(z) = 0.5z^{-2}, \\ d(z) = \dfrac{1-0.8z^{-1}+0.3z^{-2}}{1-1.1z^{-1}}-(1-0.3z^{-1}), \end{cases} \tag{6-10}$$

则(6-9)式可记为

$$a(z)\cdot e(z)+b(z)\cdot I_2(z) = d(z). \tag{6-11}$$

将(6-2)式与(6-11)式相比较,可以看出它们的相似之处:(6-2)式中 a,b,d 为整数环上已知数,而(6-11)式中 $a(z),b(z),d(z)$ 为移位算子环上的已知数;(6-2)式中 x,y 为未知整数,(6-11)式中 $e(z),I_2(z)$ 为移位算子环的未知数.

(6-1)式可依数论中不定方程知识求出它的通解:

$$\begin{cases} x = -100-4\varphi, \\ y = 200+7\varphi. \end{cases} \tag{6-12}$$

其中,φ 为任意整数.

由于整数环及移位算子环都是整环,它们有相似之处.(6-9)式或(6-11)式称为移位算子环上的不定方程.用完全类似的方法可求出(6-9)式的通解为

$$\begin{cases} e(z) = \left(0.02182z^{-2}-0.389255z^{-1}-0.9+\dfrac{0.9}{1-1.1z^{-1}}\right)-0.5z^{-2}\varphi(z), \\ I_2(z) = \left(-0.01309z^{-2}+0.2684629z^{-1}-0.126852+\dfrac{0.126852}{1-1.1z^{-1}}\right)+ \\ \qquad (1-0.8z^{-1}+0.3z^{-2})\varphi(z). \end{cases} \tag{6-13}$$

其中,$\varphi(z)$ 为移位算子环上任意常数.

不难知道,要使 $e(t)=0$,即 $e(z)$ 恒为零,这时 $\varphi(z)$ 将不属于移位算子环上的元素,即不存在控制输入使 $e(t)$ 立即为零.当

$$\varphi(z) = 2.3958z^{-1}/(1-1.1z^{-1}) \tag{6-14}$$

时,可从(6-13)式求出

$$e(z) = 0.02182z^{-2}-0.389255z^{-1}-0.9+0.9(1+1.1z^{-1}+1.1^2z^{-2}), \tag{6-15}$$

利用 Z 反变换可求出 $e(t)$.上式表明:$e(t)$ 将在二步内为零.将(6-14)式中的 $\varphi(z)$ 代入(6-13)中,可求出使误差 $e(t)$ 二步内为零的控制策略 $I_2(t+1)$.

从以上分析可以看出,控制理论离散时间系统知识、数学中的数论知识与近世代数知识、经济学知识相互交叉,古老的丢蕃图方程与现代控制论的现代频域法相互交叉.

类似地,由微分方程描述的线性连续时间系统可依拉普拉斯变换

$$f(s) = \int_0^{\infty} f(t)e^{-st}dt \tag{6-16}$$

化为移位算子环上的代数方程.

(6-8) 式与(6-16) 式所示的 Z 变换与拉普拉斯变换仅用于分析线性动态系统,对非线性动态系统来讲,可用**多重拉普拉斯变换** $h(s_1,\cdots,s_n)=\int_0^\infty\cdots\int_0^\infty h(t_1,\cdots,t_n)e^{-s_1t_1}\cdots e^{-s_nt_n}dt_1\cdots dt_n$ 进行分析,或者用**多重** Z **变换**来分析非线性离散时间系统.

以上简要介绍了代数系统理论在经济学中应用的基本概念.要深入了解经济控制论的研究方向,必须熟悉如下领域:数论、近世代数、现代频域法、多重拉普拉斯变换与 Z 变换,以及相关经济学知识.

参 考 文 献

1 (波兰)奥斯卡·兰格著.经济控制论导论.杨小凯,郁鸿胜译.北京:中国社会科学出版社,1981.

2 (罗)曼内斯库 M 著.经济控制论.赵克清译.北京:中国展望出版社,1986.

3 (美)邹至庄著.动态经济系统的分析与控制.司春林,侯先荣译.北京:友谊出版公司,1983.

4 张仲俊,王翼著.控制理论在管理科学中的应用.长沙:湖南科学技术出版社,1984.

5 龚德恩著.经济控制论概论.北京:中国人民大学出版社,1988.

6 司春林著.经济控制论.北京:中国展望出版社,1989.

7 张逸民,范崇惠编著.经济控制论.上海:同济大学出版社,1988.

8 乌家培主编.宏观经济控制论.沈阳:辽宁人民出版社,1990.

9 张金水著.经济控制论——动态经济系统分析方法与应用.北京:清华大学出版社,1999.

10 Suresh D S, Gerald L T. Optimal control theory: applications to management science. Boston: Martinus Nijhoff Publishing, 1981.

第 5 篇

精算数学

编　者　吴　岚　杨静平
审校者　胡德焜

目　　录

引　言

精算学(actuarial science)是以未来的随机事件造成的经济影响为研究对象的学科,这种影响通常是以一定量的货币形式进行刻画的.精算学要解决的主要问题是:如何定量地、精确地预测未来不确定事件的发生与否和发生的程度;分析这些事件对当前经济状况的影响.

传统的精算学主要以保险经营中的量化问题为研究对象,研究保险标的物的损失风险和相关的策略(投保风险分析、保费设计、费率结构设计和保险合同设计等),同时,在一定程度上研究对保险经营的整体控制(准备金设计、破产分析等).从保险经营的角度又可以将精算学具体分为人寿保险精算学(life actuarial science)、非寿险精算学(casualty actuarial science)两大组成部分.人寿保险精算学包括利息理论、生存模型、寿险的精算现值理论、生存年金的精算现值理论、年均衡净保费和净保费准备金等.非寿险精算学包括保费定价原理、可信度理论、风险排序、损失准备金模型和奖惩系统等内容.

精算学是一门涉及数学、统计学、经济学、金融学和保险理论等多门学科的综合学科,它的重要基础之一是数学和统计学,它们是精算学定量分析系统的核心.

1　利息理论

1.1　利息基本函数

1.1.1　累积函数

初始时刻1个货币单位的本金累积到时刻 $t(t>0)$ 的价值,称为**累积函数**(accumulation function),用 $a(t)$ 表示.它具有如下性质:

1° $a(0)=1$.

2° 一般情况下 $a(t)$ 为递增函数.

1.1.2　实际利率

实际利率(effective rate of interest)等于期初1个货币单位的本金经过一个计息期(也称利息换算期,一般为一年)产生的增量,这个增量(或称利息)是在期末支付的.通常简称为"实利率".用 i_n 表示第 n 个计息期产生的实利率,则有

$$i_n=(a(n)-a(n-1))/a(n-1)\quad n\geq 1. \tag{1-1}$$

1.1.3 单利

一个货币单位经过每个计息期产生的利息量为常数,即期初的1元投资在第一个计息期末的终值为 $1+i$ 元,在第二个计息期末的终值为 $1+2i$ 元,依此类推,那么,第 t 个计息期末的终值可用线性累积函数表示为

$$a(t)=1+it,\quad t\geqslant 0\text{ 为整数}. \tag{1-2}$$

如果利息的产生模式可以表示为以上方式,则称利息为**简单利息**(simple interest),或简称**单利**.称 i 为单利率.从实质上看,单利计算可以表述为:利息与经过的时间成正比.也可以用更严格的数学方法来定义单利,考虑如下的 $a(t)$ 函数:

$$a(t+s)=a(t)+a(s)-1,\quad t\geqslant 0,s\geqslant 0. \tag{1-3}$$

这意味着,经过时间 $t+s$ 产生的利息,等于经过时间 t 产生的利息与经过时间 s 产生的利息之和.可以证明,单利情形并不意味着有常数的实利率.实际上,如果设 i 为单利率,i_n 为第 n 个计息期的实利率,则有

$$i_n=\frac{a(n)-a(n-1)}{a(n-1)}=\frac{i}{1+i(n-1)},\quad n\geqslant 1, \tag{1-4}$$

它是 n 的递减函数.因此,单利计算隐含着递减的实利率对贷款人不利.

1.1.4 复利

如果利息的产生模式可用累积函数

$$a(t)=\prod(1+i_n),\quad t\geqslant 0\text{ 为整数} \tag{1-5}$$

表示,则称之为复利计算方式,简称**复利**(compound interest).任意时刻的复利计算的累积函数,应满足

$$a(t+s)=a(t)\times a(s),\quad t\geqslant 0,s\geqslant 0. \tag{1-6}$$

一般情况下,常考虑实利率为常数的情况,即 $i_n=i,n\geqslant 1$.这时,在一定条件下,复利计算的一般累积函数可表示为

$$a(t)=(1+i)^t,\quad t\geqslant 0. \tag{1-7}$$

1.1.5 现值

称复利方式下的

$$v=(1+i)^{-1} \tag{1-8}$$

为**贴现因子**(discount factor),其中 i 为实利率.

称单利方式下的

$$a^{-1}(t)=(1+it)^{-1}\quad(\text{其中 } i \text{ 为单利率}),$$

复利方式下的

$$a^{-1}(t)=(1+i)^{-t}=v^t\quad(\text{其中 } i \text{ 为实利率})$$

为**贴现函数**(discount function).

称 $(1+i)^t$ 为1个货币单位的本金在第 t 个计息期结束时的终值,称 v^t 为第 t 个计息期结束时的1个货币单位在0时刻的**现值**(present value).它们分别相当于复利计算的累积函数 $a(t)$ 在 t 的正、负半轴上的部分.

1.1.6 实贴现率

实贴现率(effective rate of discount)是指计息期内的利息收入与期满后的资金量的比值.一般用 d 表示.

第 n 个时间段内的实贴现率 d_n 的计算公式为

$$d_n=(a(n)-a(n-1))/a(n),\quad n\geqslant 1,\tag{1-9}$$

与常数复合利率类似,若贴现率也是常数,则简称为复合贴现(compound discount).

若相同的原始本金经过相同的计息期,将产生相同的终值

$$i=d/(1-d),$$

且

$$d=i/(1+i);$$

同时,有关系式

$$d=iv,\qquad d=1-v,\qquad i-d=id,$$

则称这个实利率和实贴现率是等价的(equivalent).

1.1.7 名义利率和名义贴现率

名义利率(nominal rate of interest)是指在给定的利息计算期内以之进行 m 次利息计算的利率.一般用记号 $i^{(m)}$($m\geqslant 1$ 的整数)表示.如"名义利率 $i^{(4)}=8\%$"或"季换算名利率 8%".

名义利率 $i^{(m)}$意味着每次利息计算的实际利率为 $i^{(m)}/m$.因此,由等价性定义,有如下实际利率 i 与名义利率 $i^{(m)}$的关系:

$$1+i=(1+i^{(m)}/m)^m,\tag{1-10}$$

也可以表示为

$$i=(1+i^{(m)}/m)^m-1,$$

或

$$i^{(m)}=m[(1+i)^{1/m}-1].$$

同样地,可以定义**名义贴现率**(nominal rate of discount)$d^{(p)}$($p\geqslant 1$ 的整数):

$$1-d=(1-d^{(p)}/p)^p,\tag{1-11}$$

上式也可以表示为

$$d=1-(1-d^{(p)}/p)^p,$$

或

$$d^{(p)}=p[1-(1-d)^{1/p}].$$

另外,名义利率与名义贴现率的关系为(m,p 可以不相同):

$$[1+i^{(m)}/m]^m=[1-d^{(p)}/p]^{-p}.\tag{1-12}$$

若 $m=p$,则有

$$[1+i^{(m)}/m]=(1-d^{(m)}/m)^{-1}.\tag{1-13}$$

另外有

$$i^{(m)}/m-d^{(m)}/m=(i^{(m)}/m)\times(d^{(m)}/m).$$

1.1.8 利息力和连续贴现率

如果累积函数 $a(t)$在时刻 t 的导数 $a'(t)$存在,则定义它在时刻 t 的利息力(the force of interest)函数 δ_t 为

$$\delta_t = a'(t)/a(t). \tag{1-14}$$

这时累积函数还可以表示为

$$a(t) = \exp\left(\int_0^t \delta_s \mathrm{d}s\right). \tag{1-15}$$

定义时刻 t 的连续贴现率函数 δ_t' 为

$$\delta_t' = -[a^{-1}(t)]'/[a^{-1}(t)], \tag{1-16}$$

显然有

$$\delta_t = \delta_t', \tag{1-17}$$

一般考虑利息力为常数的情形，即 $\delta_t = \delta$. 在此情况下有下面关于利息力 δ 与实利率 i 的关系式：

$$\exp(\delta) = 1 + i \tag{1-18}$$

或

$$\delta = \ln(1+i) = -\ln(v) = -\ln(1-d). \tag{1-19}$$

由此，自然有以下的排序：

$$d < \delta < i.$$

名义利率 $i^{(m)}$、名义贴现率 $d^{(p)}$ 与利息力 δ 的关系为

$$\exp(\delta) = [1 + i^{(m)}/m]^m = [1 - d^{(p)}/p]^{-p},$$

或

$$\delta = m\ln[1 + i^{(m)}/m] = -p\ln[1 - d^{(p)}/p].$$

自然有

$$\lim_{m\to\infty} i^{(m)} = \delta, \qquad \lim_{p\to\infty} d^{(p)} = \delta, \tag{1-20}$$

因此，有以下排序：

$$d < d^{(p)} < \delta < i^{(m)} < i. \tag{1-21}$$

1.2 年　　金

年金(annuity)原指一年支付一次的款项，现一般是指以相等的时间间隔持续按期收取的定额款项. 例如，房租、银行按揭贷款、汽车分期付款的款项和投资的利息收入等.

确定年金(annuity-certain)：指有确定起讫日期的年金. 它是无条件地进行定期的收取.

未定年金(contingent-annuity)：指无确定起讫日期的年金. 它是有条件地进行定期的收取.

付款期(payment period)：指两次年金收取之间的时间间隔. 又称付款周期.

1.2.1　基本年金

1. 单位期末年金(annuity-immediate)

在第一个收付款期末进行首次的收付款，随后依次分期进行. 每次的收付款金

额为1个货币单位,共计 n 次.所有这些年金的现值之和用记号"$a_{\overline{n}|i}$"表示.在无需特别说明利率的情况下,简记为"$a_{\overline{n}|}$".基本计算公式为

$$a_{\overline{n}|i} = v + v^2 + \cdots + v^n,$$

或

$$a_{\overline{n}|i} = [1 - v^n]/i, \tag{1-22}$$

所有收付款的终值之和用"$s_{\overline{n}|i}$"表示.其基本计算公式为

$$s_{\overline{n}|i} = 1 + (1 + i) + \cdots + (1 + i)^{n-2} + (1 + i)^{n-1},$$

或

$$s_{\overline{n}|i} = [(1 + i)^n - 1]/i, \tag{1-23}$$

$a_{\overline{n}|i}$与 $s_{\overline{n}|i}$的关系为

$$s_{\overline{n}|i} = a_{\overline{n}|i}(1 + i)^n, \tag{1-24}$$

或

$$1/a_{\overline{n}|i} = 1/s_{\overline{n}|i} + i.$$

2. 单位期初年金(annuity-due)

在合同生效时立即进行首次的收付款,随后依次分期进行.每次的收付款金额为1个货币单位,共计 n 次.用"$\ddot{a}_{\overline{n}|i}$"表示所有收付款的现值之和.基本计算公式为

$$\ddot{a}_{\overline{n}|i} = [1 - v^n]/d. \tag{1-25}$$

很容易得到期末年金与期初年金的关系:

第一组

$$\ddot{a}_{\overline{n}|i} = a_{\overline{n}|i}(1 + i), \tag{1-26}$$

和

$$\ddot{s}_{\overline{n}|i} = s_{\overline{n}|i}(1 + i); \tag{1-27}$$

第二组

$$\ddot{a}_{\overline{n}|i} = 1 + a_{\overline{n-1}|i}, \tag{1-28}$$

和

$$\ddot{s}_{\overline{n}|i} = s_{\overline{n+1}|i} - 1. \tag{1-29}$$

3. 永久年金

永久年金(perpetuity)是指年金永远按期收取下去,没有结束日期的年金.用"$a_{\overline{\infty}|i}$"表示永久单位期末年金的现值之和.

$$a_{\overline{\infty}|i} = v + v^2 + v^3 + \cdots = 1/i, \tag{1-30}$$

当然有

$$\lim_{n\to\infty} a_{\overline{n}|i} = 1/i.$$

永久年金与有限年金的关系可以表示为

$$a_{\overline{n}|i}=a_{\overline{\infty}|i}-v^n/i. \tag{1-31}$$

1.2.2 广义年金

1. 付款周期大于利息换算期(1年)的情况下

假定付款周期是利息换算期(1年)的整数倍，k 为每个付款周期内的利息换算次数，n 为年金的付款总次数 × k，i 为每个利息换算期内的实利率.

(1)单位期末年金

年金现值为

$$v^k+v^{2k}+\cdots+v^n=a_{\overline{n}|i}/s_{\overline{k}|i}. \tag{1-32}$$

年金终值为

$$(1+i)^n a_{\overline{n}|i}/s_{\overline{k}|i}=s_{\overline{n}|i}/s_{\overline{k}|i}.$$

(2)单位期初年金

年金现值为

$$a_{\overline{n}|i}/a_{\overline{k}|i}.$$

年金终值为 $s_{\overline{n}|i}/s_{\overline{k}|i}$.

2. 付款周期小于利息换算期(1年)的情况下(常见情况)

假定利息换算期(1年)是付款周期的整数倍；m 为每个利息换算期(1年)内的付款次数；n 为年金的付款总次数/m，即付款总次数为 mn；i 为每个利息换算期(1年)内的实利率.

(1)单位期末年金

在每个收付款期的期末收付款 $1/m$ 个货币单位，相当于每个利息换算期(1年)内的总收付款额为1个货币单位. 年金现值记为

$$a^{(m)}_{\overline{n}|i}=[v^{1/m}+v^{2/m}+\cdots+v^{n-1/m}+v^n]/m=\frac{1-v^n}{i^{(m)}}, \tag{1-33}$$

年金终值

$$s^{(m)}_{\overline{n}|i}=a^{(m)}_{\overline{n}|i}(1+i)^n=\frac{(1+i)^n-1}{i^{(m)}}.$$

广义年金的现值和终值与标准年金的现值和终值存在如下关系式：

$$a^{(m)}_{\overline{n}|i}=\frac{i}{i^{(m)}}a_{\overline{n}|i}, \tag{1-34}$$

$$s^{(m)}_{\overline{n}|i}=\frac{i}{i^{(m)}}s_{\overline{n}|i}.$$

(2)单位期初年金

在每个付款期的期初付款 $1/m$ 元. 年金现值记为

$$\ddot{a}_{\overline{n}|i}^{(m)}=[1+v^{1/m}+v^{2/m}+\cdots+v^{n-1/m}]/m=\frac{1-v^n}{d^{(m)}}=\frac{i}{d^{(m)}}a_{\overline{n}|i},$$

终值为

$$\ddot{s}_{\overline{n}|i}^{(m)}=\ddot{a}_{\overline{n}|i}^{(m)}(1+i)^n=\frac{(1+i)^n-1}{d^{(m)}}=\frac{i}{d^{(m)}}s_{\overline{n}|i}.$$

同时,还有如下关系式:

$$\ddot{a}_{\overline{n}|i}^{(m)}=\frac{d}{d^{(m)}}\ddot{a}_{\overline{n}|i}$$

和

$$\ddot{s}_{\overline{n}|i}^{(m)}=\frac{d}{d^{(m)}}\ddot{s}_{\overline{n}|i},$$

另外,$\ddot{a}_{\overline{n}|i}^{(m)}$相当于比$a_{\overline{n}|i}^{(m)}$提前一次付款,也就是提前 $1/m$ 个利息换算期,即期初标准年金相当于金额为$(1+i)^{\frac{1}{m}}$的期末标准年金.现值关系式为

$$\ddot{a}_{\overline{n}|i}^{(m)}=a_{\overline{n}|i}^{(m)}(1+i)^{\frac{1}{m}}=[\frac{i}{i^{(m)}}+\frac{i}{m}]a_{\overline{n}|i},\tag{1-35}$$

终值关系式为

$$\ddot{s}_{\overline{n}|i}^{(m)}=[\frac{i}{i^{(m)}}+\frac{i}{m}]s_{\overline{n}|i}.$$

3.连续年金

年金付款周期充分小,付款间隔小,付款频率快(相当于 $m\to\infty$).用"$\bar{a}_{\overline{n}|i}$"表示在$[0,n]$间任意时刻的付款额为 1,年金总时间为 n 个利息换算期,则有

$$\bar{a}_{\overline{n}|}=\int_0^n v^t\mathrm{d}t=\frac{1}{\delta}(1-\mathrm{e}^{-n\delta}).\tag{1-36}$$

另一方面,有

$$\bar{a}_{\overline{n}|}=\lim_{m\to\infty}a_{\overline{n}|}^{(m)}=\lim_{m\to\infty}\ddot{a}_{\overline{n}|}^{(m)}.$$

1.2.3 变化年金

1.年金金额等量变化

首次付 P,每次增加 Q,总计 n 次($P>0$,Q 为任意实数).如果用 A 表示这种期末年金的现值,则有

$$\begin{aligned}A&=Pv+(P+Q)v^2+(P+2Q)v^3+\cdots+[P+(n-1)Q]v^n\\&=P[1-v^n]/i+Q[a_{\overline{n}|i}-nv^n]/i\\&=Pa_{\overline{n}|i}+Q[a_{\overline{n}|i}-nv^n]/i,\end{aligned}$$

这种年金在结束时的终值为

$$Ps_{\overline{n}|i}+Q[s_{\overline{n}|i}-n]/i.\tag{1-37}$$

特别地，$P=Q=1$，称为**单位递增年金**（increasing annuity），具体年金金额为 1，2，…，n，它的现值用$(Ia)_{\overline{n}|i}$表示为

$$(Ia)_{\overline{n}|i}=a_{\overline{n}|i}+[a_{\overline{n}|i}-nv^n]/i=[\ddot{a}_{\overline{n}|i}-nv^n]/i. \tag{1-38}$$

单位递增年金在结束时的终值一般用$(Is)_{\overline{n}|i}$表示为

$$(Is)_{\overline{n}|i}=(Ia)_{\overline{n}|i}(1+i)^n=[\ddot{s}_{\overline{n}|i}-n]/i=[s_{\overline{n+1}|i}-(n+1)]/i. \tag{1-39}$$

若 $P=n, Q=-1$，称为**单位递减年金**（decreasing annuity），具体年金金额为 n，$n-1$，…，1，它的现值表示为

$$(Da)_{\overline{n}|i}=na_{\overline{n}|i}-[a_{\overline{n}|i}-nv^n]/i=[n-a_{\overline{n}|i}]/i. \tag{1-40}$$

单位递减年金在结束时的终值一般用$(Ds)_{\overline{n}|i}$表示为

$$(Ds)_{\overline{n}|i}=[n(1+i)^n-s_{\overline{n}|i}]/i. \tag{1-41}$$

2. 比例年金

一般这种期末年金的收付款方式为：首次 1 个货币单位，随后每次按比例 k 递增，总共 n 次，具体年金金额为 1，$(1+k)$，$(1+k)^2$，…，$(1+k)^n$，其现值为

$$v+(1+k)v^2+\cdots+[1+k]^{n-1}v^n=\frac{1-\left(\frac{1+k}{1+i}\right)^n}{i-k}. \tag{1-42}$$

这个公式要求 $i\neq k$. 一旦 $i=k$，就意味着利率与年金增长比例相同，相当于每次付款的现值相同，均为 v，n 次付款的现值之和为 nv.

3. 广义变化年金

问题 1：付款期 > 利息换算期（1 年），或表示为：付款期 = $k\times$ 利息换算期，n/k 表示总的付款次数（这里要求 n 是 k 的倍数，即 n 是付款总时间用利息换算期度量的结果），i 表示每个利息换算期（1 年）的实利率. 如果考虑首付 1 元，随后每次递增 1 元的方式，现值为 A，则

$$A=v^k+2v^{2k}+\cdots+nv^n/k,$$

进而有

$$A=\left[\frac{a_{\overline{n}|}}{a_{\overline{k}|}}-\frac{nv^n}{k}\right]\Big/ is_{\overline{k}|}. \tag{1-43}$$

问题 2：付款期 < 利息换算期（1 年），每个利息换算期内（1 年）付款 m 次，nm 表示总的付款次数（这里要求总的付款次数是 m 的倍数），i 表示每个利息换算期（1 年）的实利率. 考虑以下两种年金付款方式：

(1) 付款额的变化与利息换算期（1 年）的变化同步. 例如，每年内的金额不变，但金额逐年变化. 标准情形为：在前 m 次付款中（第一个利息换算期内，或第一年内），收付款额为 $1/m$，即第一年内的总金额为 1；第二个 m 次付款周期内（第二个利息换算期内，或第二年内）的收付款额为 $2/m$，即第二年内的总金额为 2；依此类推，最后一个利息换算期内的付款额为 n/m，即第 n 年内的总金额为 n. 延用前面的记号，用$(Ia)^{(m)}_{\overline{n}|i}$表示这种年金的现值，则有

$$(Ia)^{(m)}_{\overline{n}|i}=[\ddot{a}_{\overline{n}|i}-nv^n]/i^{(m)}. \tag{1-44}$$

(2)付款额的变化与付款期的变化同步.例如,每年内的收付款金额也在逐步增加.具体考虑标准方式:首次付 $1/m^2$,每次增加 $1/m^2$.第一个利息换算期(1 年)内的最后一次付款额为 $1/m$;第二个利息换算期内的最后一次付款额为 $2/m$;最后一个(第 n 个)利息换算期内的最后一次付款额为 n/m.这种年金的现值一般用 $(I^{(m)}a)_{\overline{n}|i}^{(m)}$ 表示为

$$
\begin{aligned}
(I^{(m)}a)_{\overline{n}|i}^{(m)} &= [\nu^{1/m} + 2\nu^{2/m} + \cdots + nm\nu^n]/m^2 \\
&= [\ddot{a}_{\overline{n}|i}^{(m)} - n\nu^n]/i^{(m)}.
\end{aligned} \tag{1-45}
$$

问题 3:付款金额任意变化的年金现值.

考虑一般的离散年金:设时刻 t 的付款金额为 r_t,$t = 1,2,\cdots,n$.这种年金的现值为

$$a = \sum r_t \nu^t,$$

同时,它相当于一组固定年金的和,即

$$a = \sum_{t=1}^{n} (r_t - r_{t-1}) \nu^t a_{\overline{n-t}|},$$

假设:$r_0 = 0$.

4.连续变化年金

如果用付款额函数 $f(t)$ 表示时刻 t 的付款金额,那么,用实利率 i 表示的现值为

$$a = \int_0^n f(t) \nu^t \mathrm{d}t. \tag{1-46}$$

若 $f(t) = t$,表示连续递增的年金,则现值用 $(\overline{I}\overline{a})_{\overline{n}|}$ 表示

$$(\overline{I}\overline{a})_{\overline{n}|} = \frac{\overline{a}_{\overline{n}|} - n\nu^n}{\delta},$$

另外,有

$$(I^{(m)}a)_{\overline{n}|i}^{(m)} \to (\overline{I}\overline{a})_{\overline{n}|} \quad (m \to \infty). \tag{1-47}$$

更一般地,考虑用利息力表示的连续变化年金的现值公式有

$$a = \int_0^n f(t) \exp\left(-\int_0^t \delta_s \mathrm{d}s\right) \mathrm{d}t. \tag{1-48}$$

2 寿险精算学

2.1 生存模型

2.1.1 概述

寿险以人的生命为保险标的,寿险产品的定价依据被保险人的未来的生命情

况.本章利用随机的观点来描述个体的生存规律,并对生命表做一介绍.

2.1.2 生存分布和死亡力

对于一个新生儿,设其寿命为 X,则 X 为一随机变量.记 F_X 为 X 的分布函数.假设 F_X 的密度存在,记为 f_X.

称

$$s(t)=1-F_X(t),\quad t\in[0,\infty)$$

为 X 的**生存函数**(survival function).它表示个体活过年龄 t 的可能性.分布函数和生存函数,是从不同的角度来表示生命规律的函数.

死亡力一般以 $\mu(t)$表示.分布函数 F_X 的死亡力定义为

$$\mu(t)=\frac{f_X(t)}{1-F_X(t)},\quad t\in(0,\infty).$$

$\mu(t)$可理解为一个活到年龄 t 的个体,恰在时刻 t 死亡的可能性.

2.1.3 个体的生存分布

对年龄为 x 的个体,记为(x).下面 x 均表示非负整数.记(x)的未来生存时间为 $T(x)$,即

$$T(x)=X-x.$$

假设对个体只知道其年龄 x.除此之外,不知道个体的其他任何信息,则有

$$F_{T(x)}(t)=P(X-x\leqslant t\mid X>x)=1-\frac{s(x+t)}{s(x)}.$$

又记 $T(x)$的整数部分为 $K(x)$,小数部分为 $S(x)$,则

$$T(x)=K(x)+S(x).$$

本章 2.1.2 给出了一新生儿的**死亡力**和生存分布的描述.下面具体讨论 $T(x)$ 的死亡力和生存分布.

令 $T(x)$的死亡力、分布函数、密度函数分别记为 $\mu_x(t)$, $F_{T(x)}(t)$, $f_{T(x)}(t)$.有时,$T(x)$, $K(x)$中的 x 可省略,则有

结论 2.1

1° $T(x)$分布的密度函数为

$$f_{T(x)}(t)=\frac{f(x+t)}{s(x)},\quad t\geqslant 0;$$

2° $T(x)$的死亡力函数为

$$\mu_x(t)=\mu(x+t),\quad t\geqslant 0;$$

3° $T(x)$的生存分布为

$$1-F_{T(x)}(t)=\exp\left(-\int_0^t\mu(x+s)\mathrm{d}s\right).$$

为易于表示,国际精算协会采用了一些规定,用一些具体符号表示一些特殊的概率.具体有下面几种表示:

${}_tp_x$:(x)活过 $x+t$ 的概率,即(x)至少再活 t 年的概率;

$_tq_x$:(x)在 t 年内死亡的概率;

$_{t|u}q_x$:(x)在年龄段$(x+t,x+t+u]$内死亡的概率.

另外,简记

$$_1p_x = p_x,\quad _1q_x = q_x,\quad _{t|1}q_x = _{t|}q_x.$$

2.1.4 随机生存群

设有一封闭群体,由 l_0 个新生儿组成,群体无迁出与迁入,无生育,影响群体的数目变化的唯一因素是死亡,个体的死亡相互独立.设 l_0 个个体的寿命分别为 $X_1,\cdots,X_{l_0}$.设上述每个个体的寿命服从共同的生存分布 $s(t)$.这一群体称为**随机生存群**.

设活到 x 岁的个体的期望人数为 l_x.则有

$$l_x = l_0 s(x).$$

记

$$_td_x = l_x - l_{x+t},$$

简记为

$$_1d_x = d_x,$$

则有下述结论.

结论 2.2

1° $$_tp_x = \frac{l_{x+t}}{l_x},\quad _tq_x = \frac{_td_x}{l_x};$$

2° $$\frac{\mathrm{d}l_x}{\mathrm{d}x} = -l_x\mu(x),\quad l_{x+t} = l_x\exp\left(-\int_x^{x+t}\mu(s)\mathrm{d}s\right).$$

2.1.5 生命表

生命表的指数 l_0,是指在 0 岁有 l_0 个个体.生命表是用来描述 l_0 个个体的未来规律的数据表.

对每一整数年龄 x,生命表中包含 $_tq_x,l_x,d_x,L_x,T_x,\mathring{e}_x$ 项.

$_tq_x,l_x,d_x$ 的含义在前面已经介绍过.另外的三个量中,

$$\mathring{e}_x = E(T(x)),$$

称为(x)的期望剩余寿命,其实际含义是一年龄为 x 的个体,预计的未来生存时间.

$$L_x = \int_0^1 l_{x+s}\mathrm{d}s,\quad T_x = \int_0^\infty l_{x+s}\mathrm{d}s.$$

对于初始年龄为 0 的 l_0 个个体组成的群体,T_x 表示群体在年龄 x 以后的生存的总时间的期望,L_x 表示群体在年龄 x 和年龄 $x+1$ 期间生存时间的期望.

生命表的形式如表 2-1 所示,该表摘录于美国的 1989 - 1991 年的男性生命表.$l_0 = 100\ 000$.

表2-1

年龄/岁	${}_tq_x$	l_x/人	d_x/人	L_x/h	T_x/h	$\mathring{e}_x$/岁
20~21	0.00155	97 854	151	97 778	5 211 116	53.25
21~22	0.00161	97 703	158	97 624	5 113 338	52.34
22~23	0.00167	97 545	162	97 464	5 015 714	51.42
23~24	0.00170	97 383	166	97 300	4 918 250	50.50
24~25	0.00173	97 217	168	97 133	4 820 950	49.59
25~26	0.00174	97 049	169	96 994	4 723 817	48.67
26~27	0.00176	96 880	171	96 795	4 626 853	47.76
27~28	0.00180	96 709	174	96 622	4 530 058	46.84
28~29	0.00187	96 535	180	96 445	4 433 436	45.93
29~30	0.00196	96 355	189	96 261	4 336 991	45.01
30~31	0.00205	96 166	197	96 067	4 240 730	44.10
31~32	0.00215	95 969	206	95 866	4 144 663	43.19
32~33	0.00224	95 763	215	95 655	4 048 797	42.28
33~34	0.00234	95 548	224	95 436	3 953 142	41.37
34~35	0.00245	95 324	233	95 207	3 857 706	40.47
35~36	0.00257	95 091	245	94 968	3 762 499	39.57
36~37	0.00270	94 846	255	94 719	3 667 531	38.67
37~38	0.00282	94 591	267	94 457	3 572 812	37.77
38~39	0.00293	94 324	277	94 186	3 478 355	36.88
39~40	0.00304	94 047	286	93 904	3 384 169	35.98
40~41	0.00315	93 761	295	93 613	3 290 265	35.09
41~42	0.00328	93 466	307	93 312	3 196 652	34.20
42~43	0.00344	93 159	321	92 998	3 103 340	33.31
43~44	0.00365	92 838	339	92 669	3 010 342	32.43
44~45	0.00390	92 499	360	92 319	2 917 673	31.54
45~46	0.00421	92 139	389	91 945	2 825 354	30.66
46~47	0.00457	91 750	419	91 540	2 733 409	29.79
47~48	0.00496	91 331	454	91 104	2 641 869	28.93
48~49	0.00537	90 877	488	90 634	2 550 765	28.07
49~50	0.00580	90 389	524	90 127	2 460 131	27.22

2.1.6 分数年龄段的假设

对非负整数 x,

$$T(x)=K(x)+S(x).$$

在精算中,生命表可根据相应的理论构造出来.在生命表中,只给出了整数年龄上的死亡规律.因此,由生命表只能得到 $K(x)$的分布情况,无法得到 $S(x)$的分布.但在实际应用中,也需要在非整数年龄上的分布.这时,可采取的一个方法是对分数年龄段上的生存函数做一定的假设.

具体的问题可描述为:已知 q_x,如何得到${}_tq_x$,$t\in(0,1)$?

在精算中,在年龄段$[x,x+1)$,常对生存分布采用下面三个假设:

(1)线性插值(或死亡均匀分布)假设

$$s(x+t)=(1-t)\times s(x)+t\times s(x+1),\quad t\in[0,1];$$

(2)指数插值(或死亡力常数)假设

$$\ln s(x+t)=(1-t)\times\ln s(x)+t\times\ln s(x+1),\quad t\in[0,1];$$

(3)调和插值(或 Balducci)假设

$$\frac{1}{s(x+t)}=\frac{1-t}{s(x)}+\frac{t}{s(x+1)},\quad t\in[0,1].$$

结论 2.3

1° 对年龄段$[x,x+1)$,若生存分布满足线性插值假设,则对 $t\in[0,1]$,有

$$\mu_{x+t}=\frac{q_x}{1-t\times q_x},\quad {}_tq_x=tq_x,\quad f_{T(x)}(t)=q_x;$$

2° 设对任意非负整数 y,在年龄段$[y,y+1]$,生存分布满足线性插值假设,则 $K(x)$与 $S(x)$相互独立,且 $S(x)$服从(0,1)均匀分布.

由上面的结论可知,在线性插值假设下,$K(x)$与 $S(x)$相互独立,并且 $S(x)$服从(0,1)均匀分布.这一性质,给处理一些问题带来极大的方便.因此,线性插值假设是寿险中经常采用的假设.

2.2 寿险的精算现值理论

2.2.1 概述

人寿保险以人的生命或身体为保险标的,以人的生或死为保险事故.当保险事故发生时,保险人对被保险人给付保险金额.因此,作为保险产品,有两个数量问题:一是保险合同中的保险金额,另一是保险费.保险金额的支付,是与被保险人寿命相关的.保险费的交纳分为趸交(一次性交纳)、分期交纳等情况.被保险人分期交纳的保费的数量,也是与被保险人的寿命相关的.因此,由于人的寿命的不确定性,保险公司未来是否赔付,在签定保单时,也是不定的.在分期交纳保费的条款下,每个投保人实际交纳的保费总量,也具有不确定性.如何用一种确定性的观点来描述这种不确定性,是本节考虑的主要问题.

本节主要讨论寿险中常用的给付模型.其中关键的因素有两个:一是被保险人的寿命,另一是利率水平.由于寿险产品多为长险,因此,要考虑投保人交纳的保费的时间价值.这主要通过在保险产品定价过程中,采用复利的方式来对保费计息.利率的选择依据是保险公司的投资状况.

实际上,利率是在不断地变化的.本部分中,恒假定利率为常数,不考虑利率变动的情况,以 i 表示年利率,令 $\nu=\frac{1}{1+i}$ 表示贴现因子,$\delta=\lg(1+i)$ 表示利息力,$d=\frac{i}{1+i}$ 表示贴现率,并引入**精算现值**(**APV**)的概念.未来随机给付的现值的期望,称为其精算现值.

下面就寿险的具体险种,定期死亡保险、终身寿险、n 年期生存保险来讨论精算现值.

2.2.2　定期死亡保险

定期死亡保险也称为定期寿险,是以被保险人在规定期限内发生死亡事故为前提,由保险人负责给付保险金额的人寿保险.若期限届满,被保险人仍生存,则保险人不再承担保险责任.

对被保险人(x)的 n 年期的死亡保险,若死亡发生在 n 年期内,则保险人负责给付一个单位金额.给付时刻可分为两种情况:一种为在死亡后立即给付,另一种是在死亡的保单年度末给付.

在死亡后立即给付时,现值函数

$$Z=\nu^{T(x)}I_{\{T(x)\leqslant n\}},$$

精算现值记为 $\overline{A}^{1}_{x:\overline{n}|}$.

在死亡的保单年度末给付时,现值函数

$$Z'=\nu^{K(x)+1}I_{\{T(x)\leqslant n\}},$$

精算现值记为 $A^{1}_{x:\overline{n}|}$,则有

结论 2.4

对正整数 $n\geqslant m>0$,有

$$\overline{A}^{1}_{x:\overline{n}|}=\int_0^n \nu^t{}_tp_x\,\mu_x(t)\mathrm{d}t,\qquad A^{1}_{x:\overline{n}|}=\sum_{j=0}^{n-1}\nu^{j+1}{}_jp_xq_{x+j}.$$

在死亡后立即给付和死亡的年度末给付的情况下,其精算现值有下面的关系:

结论 2.5

设对每一非负整数 y,在年龄段$[y,y+1)$,若生存分布满足线性插值假设,则

$$\overline{A}^{1}_{x,\overline{n}|}=\frac{i}{\delta}A^{1}_{x,\overline{n}|}.$$

上一结论给出了线性插值假设的具体应用.在这一假设下,对 n 年期寿险,死亡后立即给付保额和在死亡的年度末支付保额两种情况下的精算现值的比值,只与利率有关,与死亡率无关.

2.2.3 终身死亡保险

终身死亡保险又称终身寿险，是指被保险人在死亡发生后，由保险人负责给付保险金额的人寿保险.在理论上，终身寿险可以看做定期寿险的极限情况.定期寿险的一些结论，对终身寿险也成立.

对于在年龄 x 签发的终身寿险保单，若死亡保险金额为一个单位，在死亡时刻支付和死亡的保单年度末支付两种情况下，精算现值分别记为 $\overline{A}_x$，A_x，则有

结论 2.6

$$\overline{A}_x = \int_0^{\infty} \nu^t {}_t p_x \mu_x(t)\mathrm{d}t, \quad A_x = \sum_{j=0}^{\infty} \nu^{j+1} {}_j p_x q_{x+j}.$$

2.2.4 定期生存保险

定期生存保险，是指被保险人生存至保险期界满后，由保险人按保险合同的规定负责给付被保险人保险金的人寿保险.在保险期内被保险人死亡，保险人不负保险责任.

若被保险人(x)的 n 年期生存保险，保险金额为一个单位，则保险人给付的现值函数

$$Z = \nu^n I_{\{T(x) > n\}}.$$

精算现值记为 $A_{x:\overline{n|}}^{\ \ 1}$或${}_nE_x$，则

$$_nE_x = \nu^n {}_n p_x.$$

2.3 生存年金的精算现值理论

2.3.1 概述

生存年金是指在一定的期限内，若年金领取人生存则给付，若在年金支付时刻年金领取人死亡则不予支付的年金.

本章讨论两种生存年金(即定期生存年金、终身生存年金)和两种给付情况(即每年在年初给付、连续给付).

2.3.2 连续生存年金

下面考虑年金连续给付，且给付速率为 1 的情况.对于 n 年期生存年金，其现值函数为 $\overline{a}_{\overline{T(x)\wedge n|}}$，其精算现值记为 $\overline{a}_{x:\overline{n|}}$.作为终身生存年金，现值函数为 $\overline{a}_{\overline{T(x)|}}$，其精算现值记为 $\overline{a}_x$.终身生存年金，可看作定期的极限情况.

结论 2.7

$$\overline{a}_{x:\overline{n|}} = \int_0^n \nu^t {}_t p_x \mathrm{d}t, \quad \overline{a}_x = \int_0^{\infty} \nu^t {}_t p_x \mathrm{d}t.$$

生存年金与寿险有很强的联系.由利息理论中的公式

$$\bar{a}_{\overline{t|}}=\frac{1-v^t}{\delta},$$

易于得到下面的结论.

结论 2.8

$$\delta\bar{a}_x+\bar{A}_x=1.$$

对于结论 2.8,可具体解释为:若(x)个体投资一个单位的本金,投资的利息力为 δ,则通过下面的方式来支付利息和返还本金:在投资者生存期间内,连续支付利息力 δ,在投资者死亡时刻,返还一个单位的本金.

2.3.3　期初生存年金

下面考虑年金在每年年初支付,且每次给付额为1个单位的情况.对于(x)的终身生存年金和 n 年期生存年金,精算现值分别记为 $\ddot{a}_x$,$\ddot{a}_{x:\overline{n|}}$,其对应的现值函数可表示为

$$\text{终身生存年金}:\ddot{a}_{\overline{K(x)+1|}};$$

$$n\text{ 年期生存年金}:\ddot{a}_{\overline{(K(x)+1)\wedge n|}}.$$

结论 2.9

$$\ddot{a}_{x:\overline{n|}}=\sum_{j=0}^{n-1}v^j{}_jp_x,\quad \ddot{a}_x=\sum_{j=0}^{n-1}v^j{}_jp_x.$$

另外,还可得到结论 2.10.

结论 2.10

$$d\ddot{a}_x+A_x=1.$$

结论 2.10 可解释如下:若(x)投资一个单位的本金,利息率为 i,则可通过下面的方式支付利息和返还本金:在投资者生存期间内,每年支付一次利息 d,支付从投资开始进行;当投资者死亡后,在其死亡年度末返还本金.

2.4　年均衡净保费

2.4.1　概述

保费是投保人向保险人购买保险产品所支付的价格.从承保人的角度看,其必须收集足够的保费,来用于其将来的赔偿.而从被保险人的角度来看,并不希望交纳过多的保费.因此,对保险产品定价,不仅要考虑保险人的利益,同时,也要考虑投保人的情况.

在精算中,投保人投保时交纳的保费,称为毛保费.毛保费实际上由净保费和附加保费两个部分组成.

毛保费中的附加保费部分,可分为三个部分:一部分为保险中的费用部分,如营销费用,体检费用等;另一部分为考虑风险因素的风险附加,保险人在接受投保人的风险的同时,应得到一定量的风险酬金;第三部分为保险公司的利润部分,用

于满足保险公司的盈利目标.包含净保费与费用的保费部分,称为费用负荷保费.

净保费,定义其满足等式

投保人交纳的净保费的精算现值 = 保险人给付金额的精算现值,

由上述等式确定保费的准则,称之为均衡准则.

净保费是在不考虑保险人的费用、风险因素及利润的情况下,被保险人的风险成本.净保费的计算,需要在给定的利率与死亡率的条件下进行.利率是反映保险人投资收益的量化指标,死亡率是对投保人的生存规律的刻画,一般以生命表的形式给出.

本节讨论每年交纳一次保费,并且每年交纳相同的净保费的情况.其中每年交纳的净保费部分,称为**年均衡净保费**.

2.4.2 年均衡净保费

下面分死亡时刻给付和死亡年度末给付两种情况来讨论净保费.

1.年度末给付的情况

对于在 x 年龄签发的单位保额的 n 年期寿险,在保额在被保险人死亡的保单年度末给付的情况下,交费期为 n 年的年均衡净保费记为 $P^1_{x:\overline{n}|}$,则保险人签单的损失量

$$L = v^{K(x)+1} I_{\{T(x)\leqslant n\}} - P^1_{x:\overline{n}|}\ddot{a}_{\overline{(K(x)+1)\wedge n}|},$$

由 $E(L)=0$,则有

$$A^1_{x:\overline{n}|} = P^1_{x:\overline{n}|}\ddot{a}_{x:\overline{n}|},$$

即

$$P^1_{x:\overline{n}|} = \frac{A^1_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}}.$$

从上面的等式可以看出,年均衡净保费实际上是给付额的精算现值与保费交纳期的生存年金的精算现值的比值.

对于保险金额为 1 的终身寿险,其年均衡净保费记为 P_x.对于 n 年期生存保险,其年均衡净保费记为 $P_{x:\overline{n}|}^{\ \ 1}$.类似地,有

$$终身寿险:P_x = \frac{A_x}{\ddot{a}_x};$$

$$n\ 年期生存保险:P_{x:\overline{n}|}^{\ \ 1} = \frac{A_{x:\overline{n}|}^{\ \ 1}}{\ddot{a}_{x:\overline{n}|}}.$$

2.死亡时刻给付的情况

对于在 x 年龄签发的单位保额的 n 年期寿险,在保额在被保险人死亡后立即给付的情况下,年均衡净保费记为 $P(\overline{A}^1_{x:\overline{n}|})$,终身寿险的年均衡保费记为 $P(\overline{A}_x)$,则有

$$终身寿险:P(\overline{A}_x) = \frac{\overline{A}_x}{\ddot{a}_x};$$

$$n\text{ 年期寿险}:P(\overline{A}_{x:\overset{1}{\overline{n|}}})=\frac{\overline{A}_{x:\overset{1}{\overline{n|}}}}{\ddot{a}_{x:\overline{n|}}}.$$

3. 二种情况的关系

对于前面讨论的两种情况,相互之间有下面的关系:

结论 2.11

对每一非负整数 y,在年龄段$[y,y+1)$,若生存分布满足线性插值假设,则有

$$P_x=\frac{\delta}{i}P(\overline{A}_x);$$

$$P(A^1_{x:\overline{n|}})=\frac{\delta}{i}P(\overline{A}^1_{x:\overline{n|}}).$$

2.5 净保费准备金

2.5.1 概述

保单的购买,可通过一次性支付或分期支付来进行.对于死亡率来说,随着年龄的增长,死亡率会增大.因此,对于一些保单,在签单后的前几年,每年交纳的保费超过每年的保险成本.对超过前几年保险成本的资金,应该为保单持有者所拥有.这部分资金,实际上是保险公司的负债.因此,从保险人角度看,这部分资金,应该用于将来的给付,这就是准备金.

对于准备金,还可以从另一个角度来看.对于一个生效的保单,保险人面临未来的两个资金流动情况:保额的给付与投保人交纳保费.

在此时刻,

给付额的现值 - 投保人交纳保费的现值,

是保险人未来的负债.因此,保险人应该留有部分资金用于将来的给付.其期望值称为**净保费准备金**.

本节在讨论净保费准备金时,没有考虑保险中的费用、利润等其他因素.

2.5.2 完全离散模型下的净保费准备金模型

所谓完全离散模型,是指保费在每一保单年度初交纳,保额在死亡年度末给付的情况.

1. 净保费准备金的公式

考虑下面的模型:有一个在 x 岁签单的个体(x),

(1)死亡的保险金额在死亡的保单年度末支付;

(2)保费在每一保单年度年初支付;

(3)在第 j 个保险年度的死亡保险金额为 B_j, $j=1,2,\cdots$;

(4)在第 j 个保单年度的净保费交纳为 π_{j-1}, $j=1,2,\cdots$.

设 h 为非负整数.对于在 X 岁签单的个体,在 $x+h$ 时刻保险人的未来损失函数为

$$_hL = \left(B_{K(x)+1}\nu^{K(x)+1-h} - \sum_{j=h}^{K(x)}\pi_j\nu^{j-h}\right)I_{\{K(x)\geqslant h\}},$$

保险人在 h 时刻的净保费准备金 $_hV$ 定义为

$$_hV = E[_hL \mid K(x)\geqslant h].$$

于是有

命题 1

$$\begin{aligned}
_hV &= \sum_{j=0}^{\infty}\left(b_{h+j+1}\nu^{j+1} - \sum_{k=0}^{j}\pi_{h+k}\nu^{k}\right)_j p_{x+h}q_{x+h+j} \\
&= \sum_{j=0}^{\infty} b_{h+j+1}\nu^{j+1}{}_jp_{x+h}\,q_{x+h+j} - \sum_{j=0}^{\infty}\pi_{h+j}\nu^j{}_jp_{x+h}.
\end{aligned}$$

2. 递推公式

下面考虑净保费和损失函数方差的递推公式. 令

$$C_h = \nu B_{h+1}I_{\{K(x)=h\}} - \pi_h I_{\{K(x)\geqslant h\}},$$

$$\Lambda_h = \nu B_{h+1}I_{\{K(x)=h\}} + \nu_{h+1}VI_{\{K(x)\geqslant h+1\}} - (\pi_h + {}_hV)I_{\{K(x)\geqslant h\}},$$

则 C_h 表示在时刻 h, 在年度 $(h, h+1)$ 的净资金损失的现值; Λ_h 表示在时刻 h 保险人的资金损失和此年的负债变化的和.

定理 1

(1) $$_hL = \sum_{j=h}^{\infty}\nu^{j-h}C_j = C_h + \nu_{h+1}L,$$

(2) $$\pi_h + {}_hV = \nu B_{h+1}q_{x+h} + \nu_{h+1}Vp_{x+h}.$$

定理 1 (2)给出了净保费准备金的递推公式. 在第 h 年底的净保费准备金 $_hV$ 和在下一年初的保费 π_h 的和, 使得在第 $h+1$ 年底, 每个生存的人得到 $_{h+1}V$, 每个死亡的人得到 B_{h+1}.

命题 2

(1) 对 $g\leqslant h<j$,

$$\operatorname{cov}[\Lambda_h, \Lambda_j \mid K(x)\geqslant g] = 0;$$

(2) $$L_h = \sum_{j=h}^{\infty}\nu^{j-h}\Lambda_j + {}_hV;$$

(3) $$\operatorname{var}(\Lambda_h \mid K(x)\geqslant h) = [\nu(B_{h+1} - {}_{h+1}V)]^2 p_{x+h}q_{x+h}.$$

定理 2

$$\begin{aligned}
\operatorname{var}(_hL \mid K(x)\geqslant h) &= \sum_{j=h}^{\infty}\nu^{2(j-h)}\operatorname{var}[\Lambda_j \mid K(x)\geqslant h] \\
&= \operatorname{var}[\Lambda_h \mid K(x)\geqslant h] + \nu^2\operatorname{var}[_{h+1}L \mid K(x)\geqslant h+1]p_{x+h}.
\end{aligned}$$

前面的命题和定理, 在分析准备金的风险中很重要; 同时, 它给出了计算损失函数的方差的递推公式.

2.5.3 完全离散情况下的一些具体险种

对于在 x 签单的个体 (x), 若投保终身寿险, 死亡的保险金额为 1 个单位, 每

年交纳净保费 P_x，在整数时刻 k 的净保费准备金记为${}_kV_x$，则

$$B_k = 1, \quad \pi_k = P_x, \quad k = 1, \cdots.$$

因此在时刻 k 保险人的未来损失量为

$${}_kL = (v^{K(x)-k+1} - P_x\ddot{a}_{\overline{K(x)-k+1}|})I_{\{K(x)\geq k\}},$$

根据净保费准备金的定义得

$${}_kV_x = A_{x+k} - P_x\ddot{a}_{x+k}.$$

对于 n 年期死亡险，若每年交纳净保费 $P_{x:\overset{1}{\overline{n}|}}$，其在第 k 年末的净保费准备金记为${}_kV_{x:\overset{1}{\overline{n}|}}$，则可得下面公式：

$${}_kV_{x:\overset{1}{\overline{n}|}} = A_{x+k:\overset{1}{\overline{n-k}|}} - P_{x:\overset{1}{\overline{n}|}}\ddot{a}_{x+k:\overline{n-k}|}.$$

2.5.4 净保费准备金公式

在完全离散情况下，关于终身寿险的净保费准备金，有下面几个公式(其中 k 为非负整数)：

(1)保费差公式

$${}_kV_x = (P_{x+k} - P_x)\ddot{a}_{x+k};$$

(2)缴清保险公式

$${}_kV_x = (1 - \frac{P_x}{P_{x+k}})A_{x+k};$$

(3)后溯公式

$${}_kV_x = P_x\frac{\ddot{a}_{x:\overline{k}|}}{{}_kE_x} - {}_kk_x.$$

其中

$${}_kk_x = \frac{A^1_{x:\overline{k}|}}{{}_kE_x}$$

称为保险累积成本.

这三个式子，各有不同的含义.对于(x)的终身寿险来说，若每年缴纳净保费 P_x，在年龄为 $x+k$ 时，(x)仍生存，作为被保险人(x)，其未来得到的给付与其在 $x+k$ 年龄投保的终身寿险相同.对于在 $x+k$ 年龄投保的，每年交纳保费为 P_{x+k}，而(x)实际每年交纳的保费为 P_x.由此，可具体分别解释上面三式的含义如下：

(1)投保人每年少交纳 $P_{x+k} - P_x$，共计少交纳

$$(P_{x+k} - P_x)\ddot{a}_{x+k},$$

因此，保险人的${}_kV_x$ 便是这部分资金；

(2)对于个体$(x+k)$，如果每年交纳净保费 P_{x+k}，则得到精算现值 A_{x+k}，而投保人(x)实际交纳 P_x，因此，可得到的精算现值为

$$\frac{P_x}{P_{x+k}}A_{x+k}.$$

实际上,被保险人得到了 A_{x+k},其中的差额

$$(1-\frac{P_x}{P_{x+k}})A_{x+k}$$

便是保险人应负责的${}_kV_x$;

(3)被保险人投保后,在 k 年后,其交纳的保费的累积与其保险成本的差额,是被保险人的资金剩余.这部分资金,应为被保险人拥有.因此,保险人应将这部分资金留给被保险人,这便是${}_kV_x$.

2.5.5 半连续模型下的净保费准备金

所谓半连续模型,是指保费在每年年初交纳一次,保额在死亡后立即给付的情况;对于个体(x)的 n 年期寿险,在死亡后立即给付.设 $k<n$,k 年过后,个体仍生存.此时,对此个体的净保费准备金记为${}_kV(\bar{A}^1_{x:\overline{n|}})$.

保险人的未来的损失量为

$${}_kL=(v^{T(x)-k}I_{\{T(x)\leq n\}}-P(\bar{A}^1_{x:\overline{n|}})\ddot{a}_{\overline{(K(x)-k+1)\wedge(n-k)|}})I_{\{T(x)\geq k\}},$$

因此,净保费准备金为

$${}_kV(\bar{A}^1_{x:\overline{n|}})=\bar{A}^1_{x+k:\overline{n-k|}}-P(\bar{A}^1_{x:\overline{n|}})\ddot{a}_{x+k:\overline{n-k|}}.$$

结论 2.12 对于每一非负整数 y,在年龄段$[y,y+1)$,若生存分布服从线性插值假设,则

$${}_kV(\bar{A}^1_{x:\overline{n|}})=\frac{i}{\delta}{}_kV^1_{x:\overline{n|}}.$$

上面的结论,对 $n=\infty$ 仍成立.此种情况,便是终身寿险的情况.对于终身寿险,记净保费准备金为${}_kV(\bar{A}_x)$.

2.6 实例分析

设有一(20)的 3 年期寿险,死亡保险金额为 1000 元,现共有 97854 个人投保,死亡给付在年底进行.设这一群体的生存分布服从美国 1989~1991 年男性生命表(见表 2-1),即有

$$l_{20}=97854,\quad l_{21}=97703,\quad l_{22}=97545,\quad l_{23}=97383,$$

计算净保费的利率 $i=0.08$.现要求计算投保人每人每年应交纳多少净保费和保险公司资金流动情况.

1.平均净保费

由上面的数据,有

$$\begin{aligned}A^1_{20:\overline{3|}} &= \sum_{j=0}^{2}v^j{}_jp_{20}q_{20+j}\\ &= \sum_{j=0}^{2}v^j\frac{l_{20+j}}{l_{20}}\frac{l_{20+j}-l_{20+j+1}}{l_{20+j}}\end{aligned}$$

$$= \sum_{j=0}^{2} v^j \frac{l_{20+j} - l_{20+j+1}}{l_{20}}$$

$$= \frac{97854 - 97703}{97854}(1.08)^{-1} + \frac{97703 - 97545}{97854}(1.08)^{-2} + \frac{97545 - 97383}{97854}(1.08)^{-3}$$

$$= 0.004127324,$$

$$\ddot{a}_{20:\overline{3}|} = 1 + p_{20}v + {}_2p_{20}v^2$$

$$= 1 + \frac{97703}{97854}(1.08)^{-1} + \frac{97545}{97854}(1.08)^{-2}$$

$$= 2.779128661.$$

则每人每年交纳的年均衡净保费为

$$1000P^1_{20:\overline{3}|} = 1000\frac{A^1_{20:\overline{3}|}}{\ddot{a}_{20:\overline{3}|}}\text{元} = 1.48511\text{ 元}.$$

如果死亡给付在死亡后立即进行,则

$$1000P(\overline{A}^1_{20:\overline{3}|}) = \frac{i}{\delta}1000P^1_{20:\overline{3}|}\text{元} = 1.54375\text{ 元}.$$

2. 保险公司的资金流情况

假设群体完全按照生命表的规律,在21岁有97703人生存,在22岁有97545人生存,在23岁有97383人生存,则保险公司资金的流动情况如下:

第一年　共有97854人投保,每人交纳保费1.48511元,共收到保费145323.9539元,加上利息收入,则在一年末保险公司共计有156949.8703元.

此年内死亡人数151人,每人给付1000元,共给付151000元.扣除给付的,剩余5949.8703元.

在第一年末,有97703人生存.因此,对于剩余的资金,每人享有0.0609元,即是

$$1000\,{}_1V^1_{20:\overline{3}|} = 0.0609\text{ 元}.$$

第二年　年初共有97703人,每人交纳保费1.48511元,共收到145099.7023元,加上前面的剩余5949.8703元,共151049.5726元.在第二年末,资金为163133.5384元.

第二年内死亡人数158人,共计给付158000元,剩余5133.5384元.第二年末有97545人生存,故每人享有0.05263元,这便是

$$1000\,{}_2V^1_{20:\overline{3}|} = 0.05263\text{ 元}.$$

第三年　年初共有97545人,每人交纳1.48511元,共144865.0550元,加上前面的5133.5384元,共149998.5934元,加上利息,年底为161998.4808元.

此年死亡人数162人,死亡给付162000元,剩余的资金与需要给付的资金基本相同.二者的差异主要是计算的误差所致.

3 非寿险精算

3.1 保费定价原理

3.1.1 风险描述与保费定价

令 X 代表用一定货币量表示的某种经济风险,简称 X 为风险.一般要求 X 是一个非负、有有限数学期望的随机变量.

称 X 的数学期望为纯保费(pure premium).实际操作中,一般是对这个量再加上某个风险附加量,称之为净保费(net premium).净保费再加上一定的平均经营成本,称为毛保费(gross premium),即日常实际业务中所说的保费.在保险统计中主要讨论纯保费.

保费定价原理 对任意的风险 X,给出一个非负的净保费函数 $H(X)$,则称 H 为保费定价原理. H 满足的基本条件是:如果风险 X 和 Y 同分布,则有

$$H(X)=H(Y).$$

常见的三种保费定价原理是:

(1)数学期望原理

$$H_1(X)=(1+\alpha)E(X),\quad \alpha>0; \tag{3-1}$$

(2)标准差原理

$$H_2(X)=E(X)+\alpha\sqrt{\mathrm{var}(X)},\quad \alpha>0; \tag{3-2}$$

(3)方差原理

$$H_3(X)=E(X)+\alpha\mathrm{var}(X),\quad \alpha>0. \tag{3-3}$$

关于保费定价原理的主要研究问题有:

(1)保费定价原理 H 应满足的基本性质;

(2)不同定价原理之间的等价性.

3.1.2 保费定价原理的基本性质

1° 对任何风险 X 和 Y,有 $H(X+Y)\leqslant H(X)+H(Y)$.

2° 对任何风险 X 和 Y,有 $H(X)\leqslant H(X+Y)$.

3° 对任何风险 X,有 $H(X)\geqslant E(X)$.

4° 对任何风险 X,有 $P(X<H(X))<1$.

5° 对任何风险 X,有 $P(X\leqslant m)=1$,凡 $m\geqslant H(X)$.

因此,以上三种定价原理对基本性质的满足情况可以简单地用表 3-1 表示.

表3-1　三种定价原理的基本性质

基本性质	1	2	3	4
数学期望原理	+	+	+	-
标准差原理	+	-	+	-
方差原理	-	-	+	-

表中"+"表示满足性质;"-"表示不满足性质.

3.2　可信度理论

3.2.1　初期模型

设有一类保险合同,具有固定的未知风险参数 Θ,时间期限为 t 年,年索赔量用 $X_1, X_2, \cdots, X_t$ 表示.假设已知风险参数的分布函数.在给定 $\Theta=\theta$ 时,X_i 条件独立同分布.若已知某个合同的风险参数 θ,要求估计净保费以及 X_{t+1}.

定理1(Bühlmann模型)　设 $X_1, X_2, \cdots, X_t$ 为有有限方差的随机变量,给定 $\Theta=\theta$ 时,它们条件独立服从相同的分布 $F_{X|\Theta}(x;\theta)$.已知 Θ 的分布为 $U(\theta)$(一般称之为结构参数的分布).设

$$D = \{g(\cdot); g(x_1, x_2, \cdots, x_t) = c_0 + \sum_{j=1}^{t} c_j x_j, c_j \in \mathbf{R}^1, j = 1,2,\cdots,t\},$$

则有

$$\min_{g \in D} E\{[\mu(\Theta) - g(x_1, x_2, \cdots, x_t)]^2\},$$
$$\mu(\Theta) = E[X_j | \Theta = \theta]$$

的解为

$$\tilde{g}(X_1, X_2, \cdots, X_t) = z\overline{X} + (1-z)m. \tag{3-4}$$

其中

$$m = E[\mu(\Theta)], \quad \overline{X} = \frac{1}{t}\sum_{j=1}^{t} X_j, \quad z = \frac{at}{at + s^2},$$
$$a = \mathrm{var}[\mu(\Theta)] = \mathrm{var}[E(X_j | \Theta)],$$
$$s^2 = E[\sigma^2(\Theta)] = E[\mathrm{var}(X_j | \Theta)].$$

一般称上述解为最优可信度估计,称 z 为可信度因子.

定理2　若 Bühlmann 模型的结构分布为以下形式:

$$u(\theta) = q(\theta)^{-t_0} e^{-\theta x_0} / c(t_0, x_0), \quad \theta > 0;$$

X_i 的条件分布为指数族

$$f_{X|\Theta}(x, \theta) = p(x) e^{-\theta x} / q(\theta), \quad x > 0, \theta > 0,$$

其中,$p(x)$为任意非负函数,t_0 和 x_0 为正常数,$c(t_0, x_0)$为正规化系数,则有以下结论:最优可信度估计与线性可信度估计相同.即

$$E[\mu(\Theta) | X_1, X_2, \cdots, X_t] = z\overline{X} + (1-z)m.$$

3.2.2 古典模型

定理 3(古典 Bühlmann 模型) 设有 k 个合同经过 t 年的索赔记录如表 3-2 所示. X_{rj}表示第 r 年的第 j 个合同的索赔金额,第 j 个合同的结构参数为随机变量 $\Theta_j, r=1,2,\cdots,t; j=1,2,\cdots,k$.

表 3-2 k 个合同经过 t 年的索赔记录表

	Θ_1	Θ_2	$\cdots$	Θ_j	$\cdots$	Θ_k
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1k}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
t	X_{t1}	X_{t2}	$\cdots$	X_{tj}	$\cdots$	X_{tk}

表中,$(\Theta_j; X_{1j},\cdots,X_{tj})=(\Theta_j;\underline{X}_j), j=1,2,\cdots,k$.

假定$(\Theta_j;\underline{X}_j)$是独立同分布的随机向量序列,$j=1,2,\cdots,k$.对每个 j,当 $\Theta_j=\theta_j$ 固定时,$X_{1j},\cdots,X_{rj},\cdots,X_{tj}$是条件独立同分布的随机变量序列.引人记号:

$$\mu(\Theta_j)=E[X_{rj}\mid\Theta_j],\quad \sigma^2(\Theta_j)=\mathrm{var}[X_{rj}\mid\Theta_j](r=1,2,\cdots,t;j=1,2,\cdots,k);$$

$$\mathrm{cov}[X_j\mid\Theta_j]=I_{t\times t}\sigma^2(\Theta_j), m=E[\mu(\Theta_j)]=E[X_{rj}](r=1,2,\cdots,t;j=1,2,\cdots,k);$$

$$a=\mathrm{var}[\mu(\Theta_j)],\quad s^2=E[\sigma^2(\Theta_j)](j=1,2,\cdots,k).$$

则 $\mu(\Theta_j)$的最优非齐性的线性估计量为

$$\widehat{\mu}(\Theta_j)=M_j^a=(1-z)m+zM_j, \tag{3-5}$$

其中

$$M_j=\overline{X}_j=\frac{1}{t}\sum_{r=1}^{t}X_{rj};$$

$$z=\frac{at}{at+s^2}.$$

定理 4 在古典 Bühlmann 模型中,结构参数 m,a 和 s^2 的无偏估计量分别为

$$\widehat{m}=M_0=\frac{1}{k}\sum_{j=1}^{k}\overline{X}_j=\frac{1}{k}\sum_{j=1}^{k}M_j, \tag{3-6}$$

$$\widehat{s^2}=\frac{1}{k(t-1)}\sum_{j=1}^{k}\sum_{r=1}^{t}(X_{rj}-M_j)^2, \tag{3-7}$$

$$\widehat{a}=\frac{1}{(k-1)}\sum_{j=1}^{k}(M_j-M_0)^2-\frac{1}{t}\widehat{s^2}. \tag{3-8}$$

3.2.3 Bühlmann-Straub 模型

定理 5(Bühlmann-Straub 模型) 设索赔数据与古典 Bühlmann 模型相同,如表 3-3 所示.表中 W_{rj}是每次索赔的权重,$r=1,2,\cdots,t;j=1,2,\cdots,k$.

表 3-3　Bühlmann-Straub 模型索赔数据

	Θ_1	Θ_2	…	Θ_j	…	Θ_k
1	$X_{11}(W_{11})$	$X_{12}(W_{12})$	…	$X_{1j}(W_{1j})$	…	$X_{1k}(W_{1k})$
2	$X_{21}((W_{21})$	$X_{22}(W_{22})$	…	$X_{2j}(W_{2j})$	…	$X_{2k}(W_{2k})$
⋮	⋮	⋮		⋮		⋮
t	$X_{t1}(W_{t1})$	$X_{t2}(W_{t2})$	…	$X_{tj}(W_{tj})$	…	$X_{tk}(W_{tk})$

模型假设：

$$(BS_1): E[X_{rj}|\Theta_j]=\mu(\Theta_j);\quad \text{var}[X_{rj}|\Theta_j]=\sigma^2(\Theta_j)/W_{rj},$$

$$r=1,2,\cdots,t;\quad j=1,2,\cdots,k.$$

$\text{cov}[\underline{X}_{rj},\underline{X}_{qj}|\Theta_j]=0, r\neq q, r=1,2,\cdots,t;\quad q=1,2,\cdots,t;\quad j=1,2,\cdots,k.$

(BS_2)：各个合同独立，即$(\Theta_j,\underline{X}_j), j=1,2,\cdots,k$ 独立，且 $\Theta_1,\Theta_2,\cdots,\Theta_k$ 是独立同分布的，X_{rj}有有限方差.

另外，记

$$W_{\cdot j}=\sum_{r=1}^{t}W_{rj},\qquad W_{\cdot\cdot}=\sum_{j=1}^{k}W_{\cdot j},$$

$$Z_j=aW_{\cdot j}/(s^2+aW_{\cdot j}),\qquad Z_{\cdot}=\sum_{j=1}^{k}Z_j,$$

$$M_j=X_{Wj}=\sum_{r=1}^{t}\frac{W_{rj}}{W_{\cdot j}}X_{rj},\quad X_{WZ}=\sum_{j=1}^{k}\frac{Z_j}{Z_{\cdot}}X_{Wj},\quad X_{WW}=\sum_{j=1}^{k}\frac{W_{\cdot j}}{W_{\cdot\cdot}}X_{Wj},$$

则有结论：$\mu(\Theta_j)$的最优非齐性的线性估计量为

$$M_j^a=(1-Z_j)m+Z_jM_j.\tag{3-9}$$

定理 6　在 Bühlmann-Straub 模型中，结构参数 m，a 和 s^2 的无偏估计量分别为

$$\widehat{m}=M_0=X_{WZ},\tag{3-10}$$

$$\widehat{s^2}=\frac{1}{k(t-1)}\sum_{j=1}^{k}\sum_{r=1}^{t}W_{rj}(X_{rj}-X_{Wj})^2,\tag{3-11}$$

$$\widehat{a}=W_{\cdot\cdot}\frac{\sum_{j=1}^{k}W_{\cdot j}(X_{Wj}-X_{WW})^2-(k-1)\widehat{s^2}}{W_{\cdot\cdot}^2-\sum_{j=1}^{k}W_{\cdot j}^2}.\tag{3-12}$$

定理 7　在 Bühlmann-Straub 模型中，结构参数 a 的无偏和数值递推估计量(Pseudo 估计)为

$$\widehat{a}=\frac{1}{(k-1)}\sum_{j=1}^{k}Z_j(M_j-M_0)^2.\tag{3-13}$$

3.2.4　Hachemeister 回归模型

Hachemeister 考虑将“$(\Theta_j;\underline{X}_j)=\underline{X}_j|\Theta_j$ 独立同分布”的假设适当放宽，然后利用回归方法进行讨论.

定理 8(Hachemeister 回归模型)　记

$$\mu_r(\Theta_j)=E[X_{rj}\mid\Theta_j](r=1,2,\cdots,t_j;\quad j=1,2,\cdots,n)$$

或

$$\mu(\Theta_j)=E[\underline{X}_{rj}\mid\Theta_j]=(\mu_1(\Theta_j),\cdots,\mu_{t_j}(\Theta_j))^{\mathrm{T}},C_j=\mathrm{cov}[\underline{X}_j].$$

模型假设为

(H_1):合同$(\Theta_j;\underline{X}_j)$互相独立,随机变量 $\Theta_1,\Theta_2,\cdots,\Theta_n$ 是独立同分布的,

(H_2):$E[X_j\mid\Theta_j]=\boldsymbol{X}_{t_j\times n}\boldsymbol{\beta}_{n\times 1}(\Theta_j)$,

$$\mathrm{cov}[X_j\mid\Theta_j]=\sigma^2(\Theta_j)\boldsymbol{V}_j,$$

其中,$\boldsymbol{X}_{t_j\times n}$是已知的 $t_j\times n$ 阶矩阵,称之为设计矩阵;$\boldsymbol{\beta}_{n\times 1}(\Theta_j)$是未知的回归系数矩阵;$\boldsymbol{X}_{t_j\times n}$是满秩$(n<t)$矩阵,$\boldsymbol{V}_j$ 是 $t_j\times t_j$ 的权数矩阵.

结构参数

$$s^2=E[\sigma^2(\Theta_j)],$$
$$\boldsymbol{a}_{n\times n}=\mathrm{cov}[\boldsymbol{\beta}_{n\times 1}(\Theta_j)],$$
$$\boldsymbol{b}_{n\times 1}=E[\overline{\boldsymbol{\beta}}_{n\times 1}(\Theta_j)].$$

另记

$$\boldsymbol{X}=\boldsymbol{X}_{t_j\times n},$$
$$\boldsymbol{u}_j=[\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{V}_j)^{-1}\boldsymbol{X}]^{-1},$$
$$z_j=\boldsymbol{a}_{n\times n}[\boldsymbol{a}_{n\times n}+s^2\boldsymbol{u}_j]^{-1}.$$

则有结论:

1° 加权距离

$$\min_{\boldsymbol{B}_j}(\boldsymbol{X}_j-\boldsymbol{X}\boldsymbol{B}_j)^{\mathrm{T}}\boldsymbol{V}_j^{-1}(\boldsymbol{X}_j-\boldsymbol{X}\boldsymbol{B}_j),$$

极小化估计为

$$\underline{\boldsymbol{B}}_j=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_j^{-1}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_j^{-1}\underline{\boldsymbol{X}}_j=\boldsymbol{u}_j\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_j^{-1}\underline{\boldsymbol{X}}_j=(\boldsymbol{X}\boldsymbol{C}_j^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}_j^{-1}\underline{\boldsymbol{X}}_j \tag{3-14}$$
$$\boldsymbol{C}_j=s^2\boldsymbol{V}_j+\boldsymbol{X}\boldsymbol{a}_{n\times n}\boldsymbol{X}^{\mathrm{T}}.$$

2° $E[\underline{\boldsymbol{\beta}}_{n\times 1}(\Theta_j)\mid\underline{\boldsymbol{X}}_j]$的最优线性估计为

$$\underline{\boldsymbol{M}}_j=z_j\underline{\boldsymbol{B}}_j+(\boldsymbol{I}_{n\times n}-z_j)\boldsymbol{b}_{n\times 1}. \tag{3-15}$$

定理 9　Hachemeister 回归模型的结构参数估计

(1)$\boldsymbol{b}_{n\times 1}$的估计为

$$\hat{\boldsymbol{b}}_{n\times 1}=(\sum_{j=1}^{k}z_j)^{-1}\sum_{j=1}^{k}z_j\underline{\boldsymbol{B}}_j. \tag{3-16}$$

(2)若用 K 表示 $t_j>n$ 的个数,则下面的两个估计量都是 s^2 的无偏估计.

$$\hat{s}^2=\frac{1}{K}\sum_{t_j>n}\hat{s}_j^2,\quad \hat{s}_j^2=\frac{1}{t_j-n}(\underline{\boldsymbol{X}}_j-\boldsymbol{X}_j\underline{\boldsymbol{B}}_j)^{\mathrm{T}}\boldsymbol{V}_j^{-1}(\underline{\boldsymbol{X}}_j-\boldsymbol{X}_j\underline{\boldsymbol{B}}_j). \tag{3-17}$$

(3)$\boldsymbol{a}_{n\times n}$的无偏估计(Pseudo 估计)

$$\hat{\boldsymbol{a}}_{n\times n}=\frac{1}{k-1}\sum_{j=1}^{n}z_j(\underline{\boldsymbol{B}}_j-\hat{\boldsymbol{b}}_{n\times n})(\underline{\boldsymbol{B}}_j-\hat{\boldsymbol{b}}_{n\times n})^{\mathrm{T}}. \tag{3-18}$$

3.3 风险排序

3.3.1 基本概念

用非负随机变量 X 表示量化的风险.在所有风险变量组成的集合上定义一种二元关系,用“$<_p$”表示.对任意风险 X,Y 和 Z,有

(1)传递性　若 $X<_p Y, Y<_p Z$,则有 $X<_p Z$;

(2)反对称性　若 $X<_p Y, Y<_p X$,则 X 与 Y 有相同的分布,

称满足以上两个条件的关系“$<_p$”为偏序.若同时有

(3)完全性　对任意的风险 X 和 Y,必须满足以下三个条件之一:

$$X<_p Y, Y<_p X, X \text{ 与 } Y \text{ 同分布},$$

则称“$<_p$”为全序.

3.3.2 随机控制序

(1)**定义1**　风险变量 X 和 Y,若满足对所有的非降实函数 $w(x)$ 有

$$E[w(X)] \leqslant E[w(Y)],$$

则称风险 Y 随机地控制风险 X,记作“$X<_{st} Y$”.

这个偏序的背景是:如果风险决策的效用函数为单调上升函数,则决策者偏好于风险 X.

(2)随机控制与风险变量的分布函数　对于任何的风险变量 X 和 Y,有

$$X<_{st} Y \text{ 等价于 } F_Y(x) \leqslant F_X(x), \text{对任意 } x \geqslant 0.$$

(3)对偶性　风险变量 X 和 Y 满足:$X<_{st} Y$,则存在风险变量 Y' 与 Y 同分布,且满足

$$P(X \leqslant Y')=1, \quad \text{记为 } X<_1 Y';$$

反之,对任意风险变量 Y',若 $X<_1 Y'$,且 Y 与 Y' 同分布,则有 $X<_{st} Y$.

(4)判别方法　如果风险变量 X 和 Y 的分布函数 $F_X(x)$ 和 $F_Y(x)$ 都是可导函数,记它们的导函数分别为 $f_X(x)$ 和 $f_Y(x)$.如果存在 $c \geqslant 0$,使

$$f_X(x) \geqslant f_Y(x) \quad (0 \leqslant x < c),$$

$$f_X(x) \leqslant f_Y(x) \quad (c \leqslant x),$$

则有 $X<_{st} Y$.

(5)应用

二项分布 $B(n,p)$　在**随机控制序**的关系中是随 n 递增的随机变量.

泊松分布 $\pi(\lambda)$　若 $\lambda_1<\lambda_2$,则 $\pi(\lambda_1)<_{st}\pi(\lambda_2)$.

指数分布 $\exp(\lambda)$,若 $\lambda_1<\lambda_2$,则 $\exp(\lambda_1)<_{st}\exp(\lambda_2)$.

方差保费原则中,$\pi_X=E(X)+\gamma \operatorname{var}(X), \gamma>0$,对随机控制序是递减的.

3.3.3 停损序

(1)**定义2**　如果风险变量 X 和 Y 满足

$$E[(X-d)_+]\leqslant E[(Y-d)_+], \text{对任意 } d\geqslant 0,$$

则称用停损序 Y 大于 X,记作 $X<_{sl}Y$.

如果存在随机变量 D 使 $X+D$ 与 Y 同分布,而且

$$P(E[D|X]\geqslant 0)=1,$$

则称随机变量 X 比 Y 小,记作 $X<_V Y$.

如果对所有凸的不减的效用函数 $u(x)$ 成立

$$E[u(-X)]\geqslant E[u(-Y)],$$

则称从风险回避的角度看,X 小于 Y,记作 $X<_{ra}Y$.

(2)等价性　关于风险变量 X 和 Y 的以下三种排序互相等价:

$$X<_{ra}Y,\quad X<_{sl}Y,\quad X<_V Y.$$

(3)判断　如果存在实数 c,对风险变量 X 和 Y 满足

$$F_Y(x)\geqslant F_X(x)\quad (0\leqslant x<c),\quad F_Y(x)\leqslant F_X(x)\quad (c\leqslant x),$$

而且 $E(X)=E(Y)$,则有 $X<_v Y$.

如果 $E(X)=E(Y)$,而且存在对 $[0,\infty)$ 的一个分割:I_1, I_2 和 I_3 为三个不交区间,且有 $I_1\cup I_2\cup I_3=[0,\infty)$,若满足

$$\mathrm{d}F_Y(x)\geqslant \mathrm{d}F_X(x)\quad (x\in I_1\cup I_2),$$

$$\mathrm{d}F_Y(x)\leqslant \mathrm{d}F_X(x)\quad (x\in I_2),$$

则有 $X<_v Y$.

3.3.4 不变性

(1)随机变量和　如果风险变量序列 $X_1, X_2, \cdots, X_n$ 与 $Y_1, Y_2, \cdots, Y_n$ 是独立的,而且满足

$$X_i<_v Y_p,\quad i=1,2,\cdots,n,$$

则有

$$\sum X_i <_v \sum Y_i.$$

(2)随机变量的混合　如果风险变量序列 $X_1, X_2, \cdots, X_n$ 与 $Y_1, Y_2, \cdots, Y_n$ 是独立的,而且满足

$$X_i<_v Y_p,\quad i=1,2,\cdots,n,$$

设有风险变量序列 $I_1, I_2, \cdots, I_n$,均为二点分布

而且满足　$P(I_i=1)=p_i=1-P(I_i=0),\quad i=1,2,\cdots,n,$

$$\sum I_i = 1,$$

则有

$$\sum I_iX_i <_v \sum I_iY_i.$$

(3)随机变量的复合　如果风险变量序列 $X_1, X_2, \cdots$ 与 $Y_1, Y_2, \cdots$ 是独立同分布的,设 N 和 N' 是独立于 $X_1, X_2, \cdots$ 和 $Y_1, Y_2, \cdots$ 的计数随机变量,而且有

$$X_i<_v Y_p,\quad i=1,2,\cdots,l,$$

$$N<_v N',$$

则有

$$\sum_{i=1}^{N} X_i <_v \sum_{i=1}^{N'} X_i; \qquad \sum_{i=1}^{N} X_i <_v \sum_{i=1}^{N} Y_i.$$

3.4　损失准备金模型

3.4.1　背景

在某些保险险种中,有些索赔很难在索赔报告的当年结案,可能要在随后的几年内逐步赔付.对于尚未赔付的索赔损失就需要积累损失准备金,有时称这类准备金为**IBNR**(incurred but not reported)**准备金**.损失准备金的大小在很大程度上决定于对已掌握的索赔延期赔付数据的分析.

最初的IBNR准备金模型是Verbeek在1972年提出的,当时研究了带自留额的超额损失再保险情形.

一般的IBNR模型的数据为:随机变量 $X_{i,j}$ 组成的矩阵($i=1,2,\cdots,k;j=1,2,\cdots,t$),表示在某个保险年度其以往 t 年的所有已经报告的索赔实际赔付记录和未来赔付的预测,$X_{i,j}$ 代表第 i 年发生的索赔在随后的第 j 年内的索赔数据.这些随机变量可以有许多实际的背景:损失比,累计索赔指数,赔付率(总索赔与保费收入之比)等.实际上,$X_{i,j}$ 可以分为两大类:已有观测数据的变量($i+j<t+1$)和需要预测的变量($i+j>t+1$).

3.4.2　梯链(chain ladder)方法

1.基本方法

设 $X_{i,j}$ 表示累计损失指数,$k=t$.

基本假设:各列变量成比例.

如果用 $C_{j,j+1}$ 表示第 j 列与第 $j+1$ 列的比例因子,即

$$C_{j,j+1} = X_{i,j+1}/X_{i,j}, \qquad i=1,2,\cdots,t,$$

自然有 $C_{j,j+1}$ 的直接估计量

$$\widehat{C}_{j,j+1} = \frac{\sum_{i=1}^{t-j} X_{i,j+1}}{\sum_{i=1}^{t-j} X_{i,j}}, \qquad j=1,2,\cdots,t-1. \tag{3-19}$$

这时,如果用 $C_{t,\infty}$ 表示索赔发生后前 t 年总的索赔支出,与发生的时间独立,那么,它的一般估计为

$$\widehat{C}_{t,\infty} = \widehat{X}_{1,\infty}/\widehat{X}_{1,t},$$

其中

$$\widehat{X}_{1,\infty} = \widehat{X}_{1,t},$$

或保守一些,有

$$\widehat{X}_{1,\infty}=1.02\widehat{X}_{1,t}.$$

因此,得到任意两列的比例因子估计

$$\widehat{C}_{j_1,j_2}=\prod_{j=j_1}^{j_2-1}\widehat{C}_{j,j+1},\qquad j_2>j_1.\tag{3-20}$$

进而有对 X_{ij} 的下三角部分的外推估计

$$\widehat{X}_{i,j}=X_{i,t+1-i}\widehat{C}_{t+1-i,j},\quad i+j>t+1.\tag{3-21}$$

这表明:可用所在列的比例因子 $C_{t+1-i,j}$ 的估计来修正已知的最近一次的观测数据 $X_{i,t+1-i}$.

2. 广义方法

从基本方法的变量矩阵可以构造另一个变量矩阵

$$D_{i,j}=X_{i,j+1}/X_{i,j},\quad i+j<t+1.\tag{3-22}$$

依据梯链方法的假定,在 j 固定时 $D_{i,j}$ 应该是常数,记为 D_j. 因此有如下估计:

$$\widehat{X}_{ij}=X_{i,t-i+1}\prod_{s=t-i+1}^{j-1}\widehat{D}_s,\quad i+j>t+1.\tag{3-23}$$

常见的估计 D_j 的方法有

(1)线性方法

$$D_{t-1}=(D_{1,t-1}+D_{2,t-1})/2,\quad D_t=D_{1t}.\tag{3-24}$$

(2)加权平均

$$\widehat{D}_j=\sum_{i=1}^{t-j+1}W_{ij}D_{ij}\Big/\sum_{i=1}^{t-j+1}W_{ij},\tag{3-25}$$

其中, $W_{ij}=X_{ij}$, 或 $W_{ij}=i+j+1$.

(3)指数

$$X_{ij}=K_j\exp(i\beta_j),\quad i=1,2,\cdots,t-j+1,\tag{3-26}$$

$$\beta_t=\beta_{t-1}.$$

3.4.3 De Vylder 最小二乘方法(乘积模型)

令 X_i 表示第 i 年发生的索赔今后所有可能延期赔付的总和, ν_j 表示 X_{ij} 中延期 j 年赔付的比例. 所以,如果 t 充分大,有

$$\sum_{i=1}^{t}\nu_i=1.$$

最小二乘的极小化方程为

$$\sum(X_{ij}-X_i\nu_j)^2,\quad \text{这里是对所有观测到的 } X_{ij} \text{ 求和}.$$

进而得

$$X_i=\sum_j X_{ij}\nu_j\Big/\sum_j \nu_j^2,$$

$$\nu_j=\sum_s X_{ij}\nu_s\Big/\sum_s \nu_s^2,$$

由初值 $v_j = 1/t$ 迭代计算.

3.4.4　回归模型(对数线性模型)

$$\ln X_{ij} = \ln r_j + \ln\lambda_{i+j-1} + u_{ij}, \quad i+j<t+1,$$

或

$$\underline{\boldsymbol{Y}} = \boldsymbol{V}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{U}},$$

其中, u_{ij}为误差项

$$\underline{\boldsymbol{Y}} = (\ln X_{11}, \cdots, \ln X_{1t}, \cdots, \ln X_{2,t-1}, \cdots, \ln X_{t1})^{\mathrm{T}},$$

$$\underline{\boldsymbol{\beta}} = (\ln r_1, \cdots, \ln r_t, \cdots, \ln\lambda_1, \cdots, \ln\lambda_t)^{\mathrm{T}},$$

$$\underline{\boldsymbol{U}} = (u_{11}, \cdots, u_{1t}, \cdots, u_{2,t-1}, \cdots, u_{t1})^{\mathrm{T}},$$

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{I}_{t\times t} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{t\times t} \\ \boldsymbol{I}_{(t-1)\times(t-1)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{(t-1)\times(t-1)} \\ \boldsymbol{I}_{(t-j)\times(t-j)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{(t-j)\times(t-j)} \\ \boldsymbol{I}_{1\times 1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{1\times 1} \end{bmatrix},$$

"**0**"表示元素为零的矩阵.

3.4.5　自回归模型

(1)基本模型

$$X_{i,j+1} = a_{i,j} + b_{i,j}X_{i-1,j+1} + c_{i,j}X_{i,j} + u_{i,j}.$$

(2)"伦敦梯链"方法

$$X_{i,j+1} = a_j + c_jX_{i,j} + u_{i,j}.$$

可以用最小二乘方法得到 a_j 和 c_j 的估计值,进而有

$$\widehat{X}_{i,t-i+2} = \widehat{a}_{t-i+1} + \widehat{c}_{t-i+1}X_{i,t-i+1},$$

$$\widehat{X}_{i,j+1} = \widehat{a}_j + \widehat{c}_j\widehat{X}_{i,j}, \quad i>t-j+1.$$

3.4.6　可信度方法

(1)De Vylder 模型

(DV_1):向量$(\boldsymbol{\Theta}_1, \underline{\boldsymbol{X}}_1^{\mathrm{T}}), \cdots, (\boldsymbol{\Theta}_k, \underline{\boldsymbol{X}}_k^{\mathrm{T}})$独立.

(DV_2):$\underline{\boldsymbol{X}}_j$ 可以表示为

$$\underline{\boldsymbol{X}}_j = \beta(\boldsymbol{\Theta}_j)\underline{\boldsymbol{Y}}_j.$$

$\beta(\boldsymbol{\Theta}_j)$是与 j 无关的关于$\boldsymbol{\Theta}_j$ 的尺度函数,$\underline{\boldsymbol{Y}}_j$ 为未知的随机向量,$\underline{\boldsymbol{Y}}_j$ 与$\boldsymbol{\Theta}_j$ 独立.

(DV_3):$\mathrm{E}[\underline{\boldsymbol{Y}}_j] = (y_1, \cdots, y_t)^{\mathrm{T}}$ 为与 j 无关的常数向量,$j=1,\cdots,k$;另外有

$$\mathrm{cov}[\underline{\boldsymbol{Y}}_j] = (r^2/p_j)\boldsymbol{I}_{t\times t},$$

其中,r^2 为未知常数;p_j 为第j 个索赔年度的权数,$\boldsymbol{I}_{t\times t}$为 $t\times t$ 阶的单位矩阵.

(DV_4):结构变量 $\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_k$ 独立同分布.

(2)**定理 10**(De Vlyder IBNR 可信度模型)　在(DV_1)~(DV_4)的假定下,有

$\beta(\Theta_j)$的可信度估计如下：

$$\widehat{\beta}(\Theta_j)=(1-z_j)b+z_j\widehat{b}_j,$$

其中，

$$\widehat{b}_j = \sum_{q\in T_j} y_q X_{jq} \Big/ \sum_{q\in T_j} y_q^2, \quad z_j = a/(a+s_j^2 w_j),$$

$$s_j^2 = E[\sigma_j^2(\Theta_j)] = \frac{s^2}{p_j}, \quad w_j = 1\Big/\sum_{q\in T_j} y_q^2,$$

T_j 表示第 j 个发生的索赔已观测到的延期赔付的年度的脚标集合；其它常数分别为

$$b=E[\beta(\Theta_j)], \quad s^2=r^2E[\beta^2(\Theta_j)], \quad a=\mathrm{var}[\beta(\Theta_j)].$$

3.5 奖惩系统

3.5.1 背景

从 20 世纪 50 年代中期开始，在某些财产保险（特别是汽车保险的许多险种）中，保险人常考虑对投保人的保费进行适当的调整：在掌握了每个投保人一段时间内的保险记录（投保、索赔、续保等）的基础上，调整投保人下一期（年度）内的保费，一般原则是对索赔记录“较好”（索赔次数少或无索赔）的投保人适当降低保费. 称这种保费系统为无索赔优待或**奖惩系统**（bonus-mauls system），简称 BMS.

一般的 BMS 由下面三个部分组成：

(1)保费等级向量. 记为 $b=(b_1,b_2,\cdots,b_s)$. b_i 表示第 i 个等级的保费水平，$i=1,2,\cdots,s$.

(2)最初的等级 i_0.

(3)等级转移规则. 如果某个投保人在前一年的索赔次数已知，那么，决定其保费水平由原等级转移到新等级的规则，一般由保险人事先规定. 可以表示为 $T_k(i)=j$, $i=1,2,\cdots,s$; $j=1,2,\cdots,s$; $k=0,1,2,\cdots$. 当索赔次数为 k 时，投保人从等级 i 转移到等级 j. 或者定义矩阵

$$T_k=(t_{ij}^{(k)}),$$

其中：$t_{ij}^{(k)}=\delta\{T_k(i)=j\}$.

若某个投保人在一年中的索赔频率用随机变量 λ 表示，$P_k(\lambda)$表示发生 k 次索赔的概率，那么该投保人从等级 i 转移到等级 j 的概率 $P_{ij}(\lambda)$为

$$P_{ij}(\lambda) = \sum_{k=0}^{\infty} P_k(\lambda) t_{ij}^{(k)}; \quad \sum_{j=1}^{s} P_{ij}(\lambda) = 1, \quad i=1,2,\cdots,s.$$

因此，某个投保人在若干年内的保费水平就形成了一个状态空间为$\{1,2,\cdots,s\}$的随机过程.

3.5.2 BMS系统的评价

评价一个BMS系统的关键是考察该系统对索赔频率变化的灵敏程度.

(1)渐进弹性 Ioimaranta(1972年)和Vepsalainen(1972年)在假设BMS为齐性马氏链的情况下,提出了下面的BMS系统弹性的定义:

$$\eta(\lambda)=\frac{\mathrm{d}\ln P(\lambda)}{\mathrm{d}\ln\lambda}, \tag{3-27}$$

其中,λ为索赔频率变量分布函数的参数;$P(\lambda)$表示索赔频率分布参数为λ时的平均稳定保费.如果记转移概率的(极限)稳定分布为

$$\boldsymbol{a}(\lambda)=(a_1(\lambda),a_2(\lambda),\cdots,a_s(\lambda))^{\mathrm{T}},$$

则有

$$P(\lambda)=[\boldsymbol{a}(\lambda)]^{\mathrm{T}}\boldsymbol{b}.$$

如果用$P_i^{(n)}(\lambda)$表示一个新的投保人最初的等级为i经过n年的保费总支出,$i=1,2,\cdots,s$,则有

$$E[P_i^{(n)}(\lambda)]=P(\lambda)n+g_i(\lambda)+\epsilon_{i,n},$$

$\epsilon_{i,n}$以指数形式趋于零($n\to\infty$).另外有

$$g_i(\lambda)=b_i-P(\lambda)+\sum_{j=1}^{s}p_{ij}(\lambda)g_j(\lambda),\quad i=1,2,\cdots,s;$$

$$\sum_{i=1}^{s}a_i(\lambda)g_i(\lambda)=0.$$

(2)瞬间弹性 Lemaire(1985年)引入折扣因子$\beta(0<\beta<1)$的概念,定义

$$\nu_i^{(n)}(\lambda)=b_i+\beta\sum_{k=0}^{\infty}p_k(\lambda)\nu_{T_k(i)}^{(n-1)}(\lambda),$$

表示索赔频率分布参数为λ的投保人最初以等级i进入系统,经过n年的连续投保所支出的所有保费的期望值.令

$$\nu_j(\lambda)=\lim_{n\to\infty}\nu_j^{(n)}(\lambda),$$

因此,向量$\boldsymbol{\nu}(\lambda)=(\nu_1(\lambda),\nu_2(\lambda),\cdots,\nu_s(\lambda))$满足方程组

$$\nu_j(\lambda)=b_j+\beta\sum_{k=0}^{\infty}p_k(\lambda)\nu_{T_k(i)}(\lambda),$$

进而,引入瞬时弹性$\mu_i(\lambda)$的概念,

$$\mu_i(\lambda)=\frac{\mathrm{d}\ln\nu_i(\lambda)}{\mathrm{d}\ln\lambda}.$$

一般是通过求解下列方程组得到$\mu_i(\lambda)$的解:

$$\frac{\mathrm{d}\nu_i(\lambda)}{\mathrm{d}\lambda}=\beta\sum_{k=0}^{\infty}\left[\frac{\mathrm{d}p_k(\lambda)}{\mathrm{d}\lambda}\nu_{T_k(i)}(\lambda)+p_k(\lambda)\frac{\mathrm{d}\nu_{T_k(i)}(\lambda)}{\mathrm{d}\lambda}\right].$$

(3)整体弹性 在渐进情况下,整体弹性定义为

$$\eta=\int_0^{\infty}\eta(\lambda)u(\lambda)\mathrm{d}\lambda;$$

在瞬时情况下,整体弹性定义为

$$\mu_i = \int_0^\infty \mu_i(\lambda) u(\lambda) d\lambda.$$

参 考 文 献

1 Kellison. The theory of interest. 2nd ed. Illinios: Irwin Professional Pub, 1991.

2 Bowers, et al. Actuarial mathematics. 2nd ed. Illinios: Society of Actuaries, 1997.

3 Gerber H G. An Introduction to Mathematical Risk Theory. Philadelphia: S S Huebner Foundation for Insurance Education, 1980.

4 Panjer H H, Willmot G E. Insurance Risk Models. Illinois: Society of Actuaries, 1992.

5 Hans U G. Life insurance mathematics. Berlin: Springer, 1995.

6 Casualty Actuarial Society. Foundations of casualty actuarial science. Virginia: Casualty Actuarial Society. 1990.

7 Robert L B. Introduction to rate-making and loss reserving for property and casualty insurance. Connecticut: ACTEX, 1993.

8 Thomas N H. Introduction to credibility theory. Connectiut: ACTEX, 1996.

9 Jean Lemaire. Bonus-Mauls system in automobile insurance. Boston: Kluwer Academic Pub, 1995.

10 Buhlmann. Mathematical models in risk theory. New York: Springer Verlag, 1970.

·经济数学卷·

第6篇

单目标与多目标线性规划

编　者　张立卫　冯恩民
审校者　马仲蕃

目　录

引　言

线性规划问题是指目标函数和约束函数都是线性函数的数学规划问题.

线性规划问题最早是由丹齐格(George B. Dantzig)在 1947 年以前提出来的.他于 1948 年出版了《线性结构的规划》一书."线性规划"这一名称是在 1948 年夏天库普曼(T.C.Koopmans)和丹齐格首次提出的.1949 年丹齐格提出了求解线性规划问题的一个有效的方法——单纯形方法.

1979 年前苏联学者哈奇杨(L.G.Khanchian)证明了绍尔(Shor)等人提出的"椭球法"的多项式复杂度.

1984 年,印度数学家卡马卡(N.Karmarkar)提出了被称之为卡马卡算法的求解线性规划的多项式算法,使线性规划内点算法的研究成为后来数学规划研究最活跃的领域之一.

多目标规划是 20 世纪 70 年代迅速发展起来的一门新兴学科,主要研究在某些条件限制下多个数值目标如何同时达到最优的问题.1896 年和 1906 年著名法国经济学家帕雷托(V.Pareto)在经济福利理论著作中提出了多目标最优化问题及帕雷托最优化概念.现代多目标规划的正式形成始于 20 世纪 50 年代.1951 年库普曼从数量经济角度对多目标最优化问题进行了基础性研究,同年库恩(H.W.Kuhn)和塔克(A.W.Tucker)从数学规划角度给出了向量极值问题的帕雷托最优解概念,为这一学科的建立奠定了重要基础.1958 年 L.Hurwicz 把多目标规划的研究推广到一般的拓扑向量空间.1963 年 L.A.Zadeh 从控制论角度研究了多目标最优控制问题.多目标规划研究的真正兴旺发达,并正式作为数学学科的一个分支进行系统研究,是 20 世纪 70 年代以后开始的,研究内容包括多目标规划问题解的概念、解的性质、解法、对偶多目标规划、不可微多目标规划及随机多目标规划等.

本篇仅讨论单目标与多目标线性规划.该类规划的普遍性和其单纯形法、内点算法的有效性,使其已成为运筹学、经济数学、管理科学、系统分析、组合最优化等学科的基本方法.

1　线性规划问题与基本定理

1.1　线性规划标准型问题

由于各种形式的线性规划问题都可以经过初等变换转化为标准型问题,因此下面仅限于讨论如下形式的标准型问题:

$$
\text{(LP)} \qquad \begin{aligned} \min \quad & c^{\mathrm{T}} x, \\ \text{s.t.} \quad & Ax = b, \\ & x \geqslant 0, \\ & x \in \mathbf{R}^n. \end{aligned} \qquad (1\text{-}1)
$$

其中 $c \in \mathbf{R}^n, A \in \mathbf{R}^{m \cdot n}, b \in \mathbf{R}^m$.集合 $\Omega = \{x \in \mathbf{R}^n \mid Ax = b, x \geqslant 0\}$ 称为(LP)的**可行域**.为叙述方便,记 A 的第 i 列为 A_i,记指标集 $K \subseteq \{1,2,\cdots,n\}$ 的第 j 个指标为 K_j,即 $A = (A_1, A_2, \cdots, A_n), K = (K_1, K_2, \cdots, K_{|K|})$,($|K|$ 表示集合 K 的基数),记 $A_K = (A_{K_1}, A_{K_2}, \cdots, A_{K_{|K|}}), c_K^{\mathrm{T}} = (c_{K_1}, c_{K_2}, \cdots, c_{K_{|K|}}), x_K^{\mathrm{T}} = (x_{K_1}, x_{K_2}, \cdots, x_{K_{|K|}})$ 和 $b_K^{\mathrm{T}} = (b_{K_1}, b_{K_2}, \cdots, b_{K_{|K|}})$.

1.2 可行域的数学刻画

可行域是一多面凸集,对它的刻画需要引入极点、棱方向与极方向的概念.Ω 的极点即它的0维面;Ω 的棱即它的1维面;Ω 的方向即它的回收方向;如果 Ω 的方向 $d \neq 0$ 不能表示为两个不同方向的正线性组合,则称 d 是 Ω 的一个极方向.

引理1 Ω 的子集 D 是 Ω 的一个面当且仅当存在指标集 $K \subseteq \{1,2,\cdots,n\}$,使得

$$D = \Omega \cap \{x \in \mathbf{R}^n \mid x_K = 0\}.$$

定理1 $x \in \Omega$ 是 Ω 的**极点**当且仅当存在指标集 $N \subset \{1,\cdots,n\}$, $|N| = n - m, x_N = 0$ 且 A_B 是非奇异的,其中 $B = \{1,\cdots,n\} \setminus N$.

定义1 如果在 $x \in \Omega$,存在 $N \subset \{1,2,\cdots,n\}$, $|N| = n - m$,满足 $x_N = 0$ 且 A_B 是非奇异的,其中 $B = \{1,2,\cdots,n\} \setminus N$,则称 x 是 Ω 的**基本可行解**,指标集 B, N 分别称为**基指标集**与**非基指标集**;如果 $x_B > 0$,则 x 是**非退化的基本可行解**,否则 x 是**退化的基本可行解**.

定义2 设 x 是 Ω 的一个极点,$D = \{z \mid z = x + \lambda d, \lambda \in [0, \lambda_0]\}(\lambda_0 > 0)$ 是从 x 出发的一条棱;如果 $\lambda_0 < +\infty$,则 $x + \lambda_0 d$ 是与 x 相邻的一个极点;如果 $\lambda_0 = +\infty$,则 D 是一条从 x 点出发沿 d 方向的半直线,此时 Ω 无界.称 d 是关于 x 点的**棱方向**.

定理2 设 $x \in \Omega$ 是极点,$B \subset \{1,2,\cdots,n\}$ 是它的基指标集,$N = \{1,2,\cdots,n\} \setminus B$ 是非基指标集,则 $d \in \mathbf{R}^n, d \neq 0$ 是关于 x 的棱方向当且仅当存在 $j \in \{1,2,\cdots,n-m\}$,有

$$d \in \{\lambda p \mid \lambda > 0\}, \quad p_B = -A_B^{-1} A_{N_j}, \quad p_N = \widehat{e}_j$$

成立,其中 $\widehat{e}_j$ 是 $\mathbf{R}^{n-m}$ 的第 j 个单位向量.

定理3 $d \in \mathbf{R}^n, d \neq 0$ 是 Ω 的一个**极方向**的充分必要条件是存在指标集 $B \subset \{1,2,\cdots,n\}$, $|B| = m$,存在 $N_j \in N = \{1,2,\cdots,n\} \setminus B$,满足

$$d \in \{\lambda p \mid p_B = -A_B^{-1} A_{N_j}, p_N = \widehat{e}_j, \lambda \geqslant 0\}, \quad p_B \geqslant 0.$$

定理1~3分别刻画了极点、棱方向和极方向,下面的引理给出极点和极方向

存在的条件.

引理 2 若 $\Omega \neq \varnothing$,则 Ω 必有极点;若 $\Omega \neq \varnothing$ 且无界,则 Ω 必有方向、极方向.

下面的定理用极点和极方向描述了**多面凸集** Ω,从这一定理可导出线性规划的基本定理.

定理 4(多面凸集的表示定理) 设 E 是 Ω 的极点集合,P 是 Ω 的极方向集合,则 Ω 可表示为

$$\Omega = \mathrm{co}E + \mathrm{cone}P.$$

其中

$$\mathrm{co}E = \{\sum_{i\in I}\lambda_i x_i \mid x_i \in E, \lambda_i \geqslant 0\ i \in I, \sum_{i\in I}\lambda_i = 1, \mid I \mid < +\infty\},$$

$$\mathrm{cone}P = \{\sum_{i\in I}\mu_i p_i \mid p_i \in P, \mu_i \geqslant 0, i \in I, \mid I \mid < +\infty\}.$$

1.3 线性规划的基本定理

单纯形法就是基于线性规划的基本定理而设计出来的,而这一定理是线性规划的核心定理之一.

定理 5(线性规划的基本定理) 若线性规划问题(LP)有有限的**最优值**,则它必可在可行域 Ω 的某极点达到,目标函数有有限最优值的充分必要条件是,$\Omega \neq \varnothing$ 对任意的极方向 $p \in P, p^{\mathrm{T}}c \geqslant 0$.

证 设 E 是 Ω 的极点集,P 是 Ω 的极方向集. 由定理 1 与定理 3 知 $\mid E \mid < +\infty$, $\mid P \mid < +\infty$. 不妨设 $E = \{x_1, x_2, \cdots, x_k\}$, $P = \{p_1, p_2, \cdots, p_l\}$. 由定理 4 可得,对任意的 $x \in \Omega$,x 可表示为

$$x = \sum_{i=1}^{k}\lambda_i x_i + \sum_{j=1}^{l}\mu_j d_j,$$

其中

$$\sum_{i=1}^{k}\lambda_i = 1, \quad \lambda_i \geqslant 0, \quad i = 1,2,\cdots,k,$$
$$\mu_j \geqslant 0, \quad j = 1,2,\cdots,l.$$

于是(LP)问题可转化为以 $\lambda \in \mathbf{R}^k, \mu \in \mathbf{R}^l$ 为变量的如下线性规划问题:

$$\begin{aligned} \min \quad & \sum_{i=1}^{k}(c^{\mathrm{T}}x_i)\lambda_i + \sum_{j=1}^{l}(c^{\mathrm{T}}d_j)\mu_j, \\ \text{s.t.} \quad & \sum_{i=1}^{k}\lambda_i = 1, \quad \lambda_i \geqslant 0, \quad i = 1,2,\cdots,k, \\ & \mu_j \geqslant 0, \quad j = 1,2,\cdots,l. \end{aligned}$$

由于 μ 的分量可以任意大,$\mathrm{co}E$ 是有界的,(LP)问题有有限值的充分必要条件是对任意的 $d_j, j = 1,2,\cdots,l$,有 $d_j^{\mathrm{T}}c \geqslant 0$ 成立.

令
$$c^{\mathrm{T}}x_{i_0} = \min\{c^{\mathrm{T}}x_i \mid i = 1,2,\cdots,k\},$$

对任意的 $x \in \Omega$，有

$$c^{\mathrm{T}}x \geqslant \sum_{i=1}^{k}(c^{\mathrm{T}}x_i)\lambda_i \geqslant c^{\mathrm{T}}x_{i_0},$$

即在 x_{i_0} 这一极点上达到最优值. 证毕.

2 单纯形法

2.1 单纯形法的思想

线性规划的基本定理为单纯形法提供了理论依据. 单纯形法的思想是从一个极点出发沿棱方向迭代到目标函数值更小的相邻极点，并一直迭代到最优极点.

依据上述思想，首先必须给出最优极点的数学刻画，其次是解决从非最优极点出发沿哪一条棱迭代的问题，然后需确定合适的步长使下一迭代点仍然是极点，最后要针对退化极点的情形给出防止循环的策略. 以下顺次讨论这些问题.

定义 1 设 $\bar{x}$ 是 Ω 的一个极点，$B \subset \{1,2,\cdots,n\}$ 是 $\bar{x}$ 的基指标集，$N = \{1,2,\cdots,n\} \setminus B$ 是它的非基指标集，称 $r(\bar{x}) = c - A^{\mathrm{T}}A_B^{-\mathrm{T}}c_B$ 是 $\bar{x}$ 处的**相对成本向量**.

用 $r(\bar{x})$ 刻画最优极点.

定理 1 设 $\bar{x}$ 是 Ω 的一个极点，若 $r(\bar{x}) \geqslant 0$，则 $\bar{x}$ 是最优极点.

证 对任意 $x \in \Omega$，$x_B = A_B^{-1}b - A_B^{-1}A_Nx_N$，$x_N \geqslant 0$，有

$$\begin{aligned} c^{\mathrm{T}}x &= c_B^{\mathrm{T}}x_B + c_N^{\mathrm{T}}x_N \\ &= c_B^{\mathrm{T}}(A_B^{-1}b - A_B^{-1}A_Nx_N) + c_N^{\mathrm{T}}x_N \\ &= c_B^{\mathrm{T}}\bar{x}_B + r(\bar{x})_N^{\mathrm{T}}x_N \\ &= c^{\mathrm{T}}\bar{x} + r(\bar{x})_N^{\mathrm{T}}x_N \end{aligned}$$

成立. 由于 $r(\bar{x})_B = 0$，$r(\bar{x}) \geqslant 0$ 等价于 $r(\bar{x})_N \geqslant 0$，并注意到 $x_N \geqslant 0$，由 x 的任意性得 $\bar{x}$ 是最优极点. 证毕.

注 1 从凸规划的库恩 - 塔克条件可证明，当 $\bar{x}$ 是非退化基本可行解时，$r(\bar{x}) \geqslant 0$ 还是 λ 为最优极点的充分必要条件.

现在分析，若 $\bar{x}$ 不是最优的极点，应沿哪条棱迭代才能使目标函数值递减. 设 d 是下降的棱方向，由第 1 章定理 3 知，$d_B = -A_B^{-1}A_{N_j}$，$d_N = \hat{e}_j$，$j \in \{1,2,\cdots,n-m\}$，且满足 $c^{\mathrm{T}}d < 0$. 由后一条件可导出

$$\begin{aligned} c^{\mathrm{T}}d &= c_B^{\mathrm{T}}d_B + c_N^{\mathrm{T}}d_N \\ &= -c_B^{\mathrm{T}}A_B^{-\mathrm{T}}A_{N_j} + c_{N_j} \\ &= r_{N_j} < 0, \end{aligned}$$

可见要选取 $j^* \in \{1,2,\cdots,n-m\}$ 满足 $r_{N_{j^*}} < 0$，它对应的棱方向 d 由 $d_B =$

$-A_B^{-1}A_{N_{j^*}}$，$d_N = \widehat{e}_{j^*}$ 来定义，是 $\bar{x}$ 处下降的棱方向.

选取了下降的棱方向后，进而选取步长．分下述三种情况：

情况 1　如果 $d \geqslant 0$（或 $d_B \geqslant 0$），$t = +\infty$ 取为步长，此时 d 是极方向，(LP)没有有限的最优值，也没有有限最优解.

情况 2　如果 $d \ngeqslant 0$，$\min\{-\bar{x}_{B_i}/d_{B_i} \mid d_{B_i} < 0\} = 0$，此时，$t = 0$ 取为步长，$\bar{x}$ 是退化的极点．关于这种情况将在 2.4 节讨论.

情况 3　如果 $d \ngeqslant 0$，$t^* = \min\{-\bar{x}_{B_i}/d_{B_i} \mid d_{B_i} < 0\} > 0$，则步长取为 $t = t^*$，下一个迭代点为

$$x' = \bar{x} + td. \tag{2-1}$$

下面的定理表明，在情况 3 时，由(2-1)式确定的 x' 是极点.

定理 2　如果 $d \ngeqslant 0$，$t^* = \min\{-\bar{x}_{B_i}/d_{B_i} \mid d_{B_i} < 0\} > 0$，则 $x' = \bar{x} + t^* d$ 是极点，它对应的基指标集 B' 与非基指标集 N' 可分别选取为

$$B' = (B_1, \cdots, B_{i^*-1}, N_{j^*}, B_{i^*+1}, \cdots, B_m), \tag{2-2}$$

$$N' = (N_1, \cdots, N_{j^*-1}, B_{i^*}, N_{j^*+1}, \cdots, N_{n-m}), \tag{2-3}$$

且 $c^{\mathrm{T}}x' < c^{\mathrm{T}}\bar{x}$，其中 i^* 满足 $1 \leqslant i^* \leqslant m$，

$$-\bar{x}_{B_{i^*}}/d_{B_{i^*}} = t^*.$$

2.2　迭代形式的单纯形方法

基于上一节的分析，本节将给出单纯形法的迭代形式，并举一简单的例子.

单纯形法(迭代形式)的步骤如下：

步 0(初始步)　求解 Ω 的极点 x_0，它对应的基指标集与非基指标集分别记为 B^0，N^0，置 $k = 0$.

步 1(最优性检验)　计算 x_k 点的相对成本向量

$$r_k = c - A^{\mathrm{T}}A^{-\mathrm{T}}_{B^k}c_{B^k},$$

若 $r_k \geqslant 0$，则 x_k 是最优极点，否则(按布兰德原则)选取 j^* 满足 $(r_k)_{N^k_{j^*}} < 0$，转步 2.

步 2(计算搜索方向)　计算 d_k

$$(d_k)_{B^k} = -A^{-1}_{B^k}A_{N^k_{j^*}}, (d_k)_{N^k} = \widehat{e}_{j^*}^k, \text{转步 3.}$$

步 3(计算步长)　若 $(d_k)_{B^k} \geqslant 0$，则问题(LP)无有限最优值和最优解，停止计算；否则(按布兰德法则)选取 i^* 满足

$$t_k = -(x_k)_{B^k_{i^*}}/(d_k)_{B^k_{i^*}} = \min\{-(x_k)_{B^k_i}/(d_k)_{B^k_i} \mid (d_k)_{B^k_i} < 0\}.$$

步 4(计算下一极点)　计算

$$x_{k+1} = x_k + t_k d_k,$$

置 $B^{k+1} = B^k \setminus \{B^k_{i^*}\} \cup \{N^k_{j^*}\}$，$N^{k+1} = N^k \setminus \{N^k_{j^*}\} \cup \{B^k_{i^*}\}$.

步 5　置 $k = k + 1$　转步 1.

注2 为了算法的完整性,进基(即选取 j^*)与离基(即选取 i^*)的原则采用布兰德原则,可以避免退化情形循环地发生. 布兰德原则将在2.4节介绍. B^{k+1} 与 N^{k+1} 的顺序可以按(2-2)式与(2-3)式排列,也可有其他定义方法,如鲁恩伯杰(Luenberger)(1973)的方法.

结合将在2.4节中叙述的布兰德原则,可以得到如下的关于单纯法的收敛性定理.

定理3 从某一极点 x_0 出发的结合布兰德进基、离基原则的单纯形法,经有限次迭代,即可得到最优极点或判断出线性规划问题无有限最优解.

注3 在单纯形法的步0阶段,要求事先知道初始的极点 x_0,这往往要解另外一个初始极点已知的线性规划问题,这即是所谓的两阶段法的第一阶段.

下面给出一简单的例子说明单纯形法的迭代过程.

例1 求解(LP)问题,其中

$$A = \begin{bmatrix} 1 & -1 & 2 & 1 & 0 \\ 4 & 2 & -1 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

$$c = (-2, -3, -1, 0, 0)^{\mathrm{T}}.$$

解 迭代0.

步0 $x_0 = (0,0,0,1,2)^{\mathrm{T}}$, $B^0 = (4,5)$, $N^0 = (1,2,3)$,

$c_{B^0} = (0,0)^{\mathrm{T}}$, $c_{N^0} = (-2,-3,-1)^{\mathrm{T}}$,

$$A_{B^0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_{N^0} = \begin{bmatrix} 1 & -1 & 2 \\ 4 & 2 & -1 \end{bmatrix}.$$

步1 最优性检验

$$(r_0)_{N^0} = c_{N^0}^{\mathrm{T}} - c_{B^0}^{\mathrm{T}} A^{-1}_{B^0} A_{N^0} = (-2, -3, -1)^{\mathrm{T}},$$

取 $j^* = 2$,即 $N^0_{j^*} = 2$.

步2 计算搜索方向 d_0

$$(d_0)_{B^0} = ((d_0)_4, (d_0)_5)^{\mathrm{T}} = -A^{-1}_{B^0} A_{N^0_2} = (1, -2)^{\mathrm{T}},$$

$$(d_0)_{N^0} = \widehat{e}_{j^*} = \widehat{e}_2 = (0,1,0)^{\mathrm{T}}.$$

步3 计算步长

$$t_0 = \min\{-(x_0)_{B^0_k}/(d_0)_{B^0_k} \mid (d_0)_{B^0_k} < 0\} = -(x_0)_{B^0_2}/(d_0)_{B^0_2} = 1,$$

取 $i^* = 2, B^0_{i^*} = 5$.

步4 计算 x_1

$$x_1 = x_0 + t_0 d_0 = (0,1,0,2,0)^{\mathrm{T}},$$

$$B^1 = (B^0_1, N^0_2) = (4,2),$$

$$N^1 = (N^0_1, B^0_2, N^0_3) = (1,5,3).$$

迭代1

步1 最优性检验

$$(r_1)^{\mathrm{T}}_{N^1} = c_{N^1} - c^{\mathrm{T}}_{B^1} A^{-1}_{B^1} A_{N^1} = (4, \frac{3}{2}, -\frac{5}{2}),$$

取 $j^* = 3$,有 $N_{j^*}^1 = 3$.

步 2 计算搜索方向 $\boldsymbol{d}_1$

$$(\boldsymbol{d}_1)_{B^1} = ((d_1)_4,(d_1)_2)^{\mathrm{T}} = -\boldsymbol{A}_{B^1}^{-1}\boldsymbol{A}_3 = (-\frac{3}{2},\frac{1}{2})^{\mathrm{T}},$$

$$(\boldsymbol{d}_1)_{N^1} = ((d_1)_1,(d_1)_5,(d_1)_3)^{\mathrm{T}} = \widehat{\boldsymbol{e}}_{j^*} = \widehat{\boldsymbol{e}}_3 = (0,0,1)^{\mathrm{T}}.$$

步 3 计算步长 t_1

$$\begin{aligned} t_1 &= \min\{-(x_1)_{B_k^1}/(d_1)_{B_k^1} \mid (d_1)_{B_k^1} < 0\} \\ &= -(x_1)_{B_1^1}/(d_1)_{B_1^1} = 4/3. \end{aligned}$$

步 4 计算 $\boldsymbol{x}_2$

$$\begin{aligned} \boldsymbol{x}_2 &= \boldsymbol{x}_1 + t_1\boldsymbol{d}_1 = (0,\frac{5}{3},\frac{4}{3},0,0)^{\mathrm{T}}, \\ B^2 &= (N_3^1,B_2^1) = (3,2), \\ N^2 &= (N_1^1,N_2^1,B_1^1) = (1,5,4). \end{aligned}$$

迭代 2

步 1 最优性检验

$$(\boldsymbol{r}_2)_{N_2}^{\mathrm{T}} = \boldsymbol{c}_{N^2} - \boldsymbol{c}_{B^2}^{\mathrm{T}}\boldsymbol{A}_{B^2}^{-1}\boldsymbol{A}_{N^2} = (9,\frac{7}{3},\frac{5}{3}) \geqslant 0,$$

得到的 $\boldsymbol{x}_2 = (0,\frac{5}{3},\frac{4}{3},0,0)^{\mathrm{T}}$ 即为最优极点.

2.3 表格形式的单纯形方法

这一节介绍单纯形方法的表格形式.设 $\boldsymbol{x}$ 是 Ω 的一个极点,对应的基指标集与非基指标集分别为 B 与 N,单纯形表格可视为如下形式的矩阵 $\boldsymbol{M}(\boldsymbol{x},B) \in \mathbf{R}^{m+1,n+1}$:

$$\boldsymbol{M}(\boldsymbol{x},B) = \begin{bmatrix} \boldsymbol{A}_B^{-1}\boldsymbol{A} & \boldsymbol{A}_B^{-1}\boldsymbol{b} \\ \boldsymbol{c}^{\mathrm{T}} - \boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{A}_B^{-1}\boldsymbol{A} & -\boldsymbol{c}_B^{\mathrm{T}}\boldsymbol{A}_B^{-1}\boldsymbol{b} \end{bmatrix}. \tag{2-4}$$

矩阵 $\boldsymbol{M}(\boldsymbol{x},B)$ 是与$(\boldsymbol{x},B,N)$一一对应的.表格形式的单纯形法就是基于对 $\boldsymbol{M}$ 的初等行变换来实现从非优极点到最优极点之迭代的.假定初始极点 $\boldsymbol{x}$ 及对应的基指标集 B 是已知的.

单纯形法(表格形式)的步骤如下:

步 1(最优性检验) 如果 $\boldsymbol{M}(\boldsymbol{x},B)$的第 $m+1$ 行中的前 n 个元素均非负,则停止计算,$\boldsymbol{x}$ 即为最优极点;否则转步 2.

步 2(选取进基指标) (按某一原则,如布兰德原则) 选取指标 $k \in \{1,2,\cdots,n\}$ 满足 $M_{m+1,k} < 0$,则 k 为进基指标,转步 3.

步 3(判断(LP) 是否有界) 若 $M_k \leqslant 0$,则停止计算,问题(LP) 没有有限最优值和最优解;否则,转步 4.

步 4(选取离基指标) 计算

$$t = \min\{M_{i,n+1}/M_{i,k} \mid M_{i,k} > 0\}$$

(按某种原则,如布兰德原则)选取 $l \in \{1,2,\cdots,m\}$,满足

$$M_{l,n+1}/M_{l,k} = t,$$

则 B_l 为离基指标.

步5(产生新的单纯形表)　以 $M_{l,k}$ 为主元,利用高斯(Gauss)变换将 M_k 变为 $\tilde{e}_l \in \mathbf{R}^{m+1}$(其中 $\tilde{e}_l \in \mathbf{R}^{m+1}$ 是第 l 个单位向量),即将 $\boldsymbol{M}(\boldsymbol{x},\boldsymbol{B})$ 左乘一$(m+1)\times(m+1)$ 的高斯变换阵,将它的第 k 列变为 $\tilde{e}_l$.记 $\boldsymbol{M}(\boldsymbol{x},\boldsymbol{B})$ 经过变换后为 $\boldsymbol{M}'$.

步6　记

$$B' = (B_1,\cdots,B_{l-1},k,B_{l+1},\cdots,B_m),$$
$$x'_{B'_i} = M'_{i,n+1}, i = 1,2,\cdots,m,$$
$$x'_j = 0, j \in \{1,2,\cdots,n\} \setminus B'.$$

置 $\boldsymbol{M}(\boldsymbol{x}',\boldsymbol{B}') = \boldsymbol{M}', \boldsymbol{x} = \boldsymbol{x}', \boldsymbol{B} = \boldsymbol{B}'$,转步1.

下面用表格形式的单纯形法求解例1.

迭代0

$$\boldsymbol{x} = (0,0,0,1,2)^{\mathrm{T}}, \quad \boldsymbol{B} = (4,5),$$

$$\boldsymbol{M}(\boldsymbol{x},\boldsymbol{B}) = \begin{bmatrix} 1 & -1 & 2 & 1 & 0 & 1 \\ 4 & 2 & -1 & 0 & 1 & 2 \\ -2 & -3 & -1 & 0 & 0 & 0 \end{bmatrix}.$$

步1　显然 $\boldsymbol{x}$ 不是最优极点.

步2　取 $k = 2$.

步3　不能判断问题(LP)无界.

步4　$l = 2, B_l = 5$.

步5　将 $\boldsymbol{M}(\boldsymbol{x},\boldsymbol{B})$ 的第2列变为 $\tilde{e}_2$ 得

$$\boldsymbol{M}' = \begin{bmatrix} 3 & 0 & \frac{3}{2} & 1 & \frac{1}{2} & 2 \\ 2 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} & 1 \\ 4 & 0 & -\frac{5}{2} & 0 & \frac{3}{2} & 3 \end{bmatrix}.$$

步6　$\boldsymbol{x}' = (0,1,0,2,0)^{\mathrm{T}}, \boldsymbol{B}' = (4,2)$.

迭代1　由 $\boldsymbol{x} = (0,1,0,2,0)^{\mathrm{T}}, \boldsymbol{B} = (4,2)$.容易得

$$\boldsymbol{M}' = \begin{bmatrix} 2 & 0 & 1 & \frac{2}{3} & \frac{1}{3} & \frac{4}{3} \\ 3 & 1 & 0 & \frac{1}{3} & \frac{2}{3} & \frac{5}{3} \\ 9 & 0 & 0 & \frac{5}{3} & \frac{7}{3} & \frac{19}{3} \end{bmatrix}.$$

由于 M' 的最后一行前 5 个元素非负,得$(0,\frac{5}{3},\frac{4}{3},0,0)^T$即最优极点.

注 4 容易证明由表格形式的单纯形法中的第5步得到的 M' 即$M(x',B')$.

2.4 布兰德原则

当线性规划的可行域有退化极点时,迭代时可能发生循环.下面举索娄(Solow)(1984)的例子加以说明.

$$\begin{aligned}
\min\quad & -2x_3-2x_4+8x_5+2x_6,\\
\text{s.t.}\quad & x_1-7x_3-3x_4+7x_5+2x_6=0,\\
& x_2+2x_3+x_4-3x_5-x_6=0,\\
& x_i\geqslant 0,i=1,2,\cdots,6.
\end{aligned}$$

初始点 $x_0=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^0=(1,2),N^0=(3,4,5,6)$.采用单纯形法可能得到下面的迭代

$$\begin{aligned}
x_1&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^1=(1,3),N^1=(2,4,5,6);\\
x_2&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^2=(4,3),N^2=(2,1,5,6);\\
x_3&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^3=(4,5),N^3=(2,1,3,6);\\
x_4&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^4=(6,5),N^4=(2,1,3,4);\\
x_5&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^5=(6,2),N^5=(5,1,3,4);\\
x_6&=(0,0,\cdots,0)^T\in\mathbf{R}^6,B^6=(1,2),N^6=(5,6,3,4).
\end{aligned}$$

可见,$(x_0,B^0,N^0)=(x_6,B^6,N^6)$,即发生了循环.

有多种避免循环发生的方法,如青恩斯方法,丹齐格,奥典和沃尔夫方法及布兰德方法.本节只介绍布兰德原则.

布兰德原则 设 $x\in\Omega$ 是单纯形法中迭代到的一个极点,它不是最优的,B,N 是 x 的基指标集和非基指标集,选取

$$j^*=\arg\min\{N_j\mid r_{N_j}<0,j=1,2,\cdots,n-m\},$$

棱方向 $d\in\mathbf{R}^n$ 定义为$d_B=-A_B^{-1}A_{N_{j^*}}$,$d_N=\widehat{e}_{j^*}$,选取

$$i^*=\arg\min\{B_i\mid -x_{B_i}/d_{B_i}=\min\{-x_{B_j}/d_{B_j}\mid d_{B_j}<0\},i=1,2,\cdots,m\},$$

则极点 $x'=x-(x_{B_{i^*}}/d_{B_{i^*}})d$ 的基指标集 B' 和非基指标集 N' 分别定义为

$$B'=(B_1,\cdots,B_{i^*-1},N_{j^*},B_{i^*+1},\cdots,B_m),$$
$$N'=(N_1,\cdots,N_{j^*-1},B_{i^*},N_{j^*+1},\cdots,N_{n-m}).$$

定理 4 按布兰德原则选取进基指标和离基指标的单纯形法不会发生循环.

证 假设按布兰德原则选取进基指标和离基指标的单纯形法在某个极点 x 处发生循环,即 x 的指标集对$[B^1,N^1]$经过 $p(p>1)$ 次迭代得到$[B^{p+1},N^{p+1}]=[B^1,N^1]$.由于不同极点不会有相同的基指标集,必有$[B^i,N^i](i=1,2,\cdots,p)$均是极点 x 的基指标集与非基指标集对,即循环中每次迭代的步长都是 0.

定义指标集 $I=\{1\leqslant i\leqslant n\mid i$ 在循环中的某次迭代由非基指标集进入基指标集,在另外的迭代由基指标集进入非基指标集$\}$,

$$i^*=\max_{i\in I} i.$$

设 i^* 在循环的某一次迭代由 N 进入 B,在之后的另一次迭代 i^* 从 B' 进入 N'. 首先有 $r_i^*<0(r=c-A^{\mathrm T}A_B^{-\mathrm T}c_B)$. 设 i^* 从 B' 进入 N' 时, d' 是迭代的搜索棱方向, 则 $d'_{i^*}<0, Ad'=0, c^{\mathrm T}d'<0$, 得到 $r^{\mathrm T}d'=c^{\mathrm T}d'<0$. 注意到 $r_{i^*}<0, d'_{i^*}<0$, 则必有 $i\in\{1,2,\cdots,n\}$ 存在, 满足 $r_id'_i<0$, 显然有 $i\neq i^*$.

下面要证明 $i\in I'$. 由 $d'_i\neq 0$, 有 i 在 B' 中, 或在 N' 中, 但它即将进入到基指标集中. 如果 $i\in B'$, 由 $i\in B^1$, 注意到 $i\in N$, 则有 $i\in I'$; 如果 $i\in N^1$, 由 $i\in N^{p+1}$, 则 $i\in I'$. 如果 i 在 N' 中, 它即将进入到基指标集中, 若 $i\in B^1$, 同样由 $i\in N$, 必有 $i\in I'$; 若此时 $i\in N^1$, 有 $i\in N^{p+1}$, 则 $i\in I'$.

再证 $i\in B'$. 因为若 $i\in N'$, 必有 $d'_i>0, r_i<0$, 由 $i\in I', i<i^*$, 得到与布兰德原则相矛盾的结果. 因此必有 $i\in B'$.

由 $i<i^*$, 有 $r_i>0$, 则 $d'_i<0$. 注意到 $x_{I'}=0$, 有 $x_i/(-d_i)=x_{i^*}/(-d_{i^*})=0$, 而 $i<i^*$, 这表明按布兰德原则应不会是 i^* 由 B' 进入 N', 这导致矛盾. 证毕.

3 线性规划的对偶理论

3.1 对偶线性规划

根据第2章定理1, x 是最优极点(B 与 N 是它的基指标集与非基指标集), 当且仅当 $r=c-A^{\mathrm T}A_B^{-\mathrm T}c_B\geqslant 0$ 时. 令 $w=A_B^{-\mathrm T}c_B$, 则这一不等式等价为

$$A^{\mathrm T}w\leqslant \overline{c}.$$

注意到最优极点 x 满足 $x_N=0$, 由此得到

$$b^{\mathrm T}w=c_B^{\mathrm T}A_B^{-1}b=c^{\mathrm T}x,$$

但通常只有

$$b^{\mathrm T}w=w^{\mathrm T}Ax\leqslant c^{\mathrm T}x;\quad Ax=b, x\geqslant 0.$$

因此从变量 w 的角度来看, 可以定义一个相关的线性规划问题

(DLP)
$$\begin{cases}\max & b^{\mathrm T}w,\\ \text{s.t.} & A^{\mathrm T}w\leqslant c.\end{cases}\tag{3-1}$$

称(3-1)式(DLP)问题是(LP)问题的**对偶问题**, w 为**对偶变量**.

例1 给出下述问题的对偶规划问题

$$\begin{cases}\min & c^{\mathrm T}x,\\ \text{s.t.} & A^{\mathrm T}x\geqslant b, x\geqslant 0.\end{cases}$$

解 对偶问题为

$$\begin{aligned} \max \quad & b^T w, \\ \text{s.t.} \quad & A^T w \leqslant c, w \geqslant 0. \end{aligned}$$

3.2 对偶理论

引理 1 给出一个原线性规划问题,其对偶问题的对偶是它自身.

定理 1(线性规划的对偶定理) 若 x_0 是一个原始可行解,以及 w_0 是对偶可行解,那么 $c^T x_0 \geqslant b^T w_0$.

定理 2 若 x_0 是原始可行的,w_0 是对偶可行的,且 $c^T x_0 = b^T w_0$,则 x_0 和 w_0 分别是问题(LP)与问题(DLP)的解;若原问题是没有下界的,则对偶问题不可行;若对偶问题没有上界,则原问题不可行.

定理 3(线性规划的强对偶定理) 考虑问题(LP)及其对偶问题(DLP),有如下关系成立:

(1) 如果原问题和对偶问题中任何一个有有限的最优解,则另一个也有有限的最优解,且它们的目标函数最优值相等;

(2) 如果任何一个问题的目标函数值无界,则另一个无可行解.

事实上,原问题与对偶问题的最优性条件是重合的,对偶问题是以原问题的乘子变量为自变量的线性规划.

定理 4(补偿松弛定理) x 与 w 分别是(LP)问题与(DLP)问题的可行解,则 x 与 w 分别为原问题与对偶问题最优解的充分必要条件是

$$x_i(c_i - e_i^T A^T w) = 0, \quad i = 1, 2, \cdots, n.$$

线性规划的最优性条件可用定理 5 表述.

定理 5(最优性条件) 给定标准型线性规划问题(LP),当且仅当存在向量 $w \in \mathbf{R}^m, r \in \mathbf{R}^n$ 满足

1° $Ax = b, x \geqslant 0$,

2° $A^T w + r = c, r \geqslant 0$,

3° $r^T x = 0$

时,向量 x 是(LP)的最优解,此时 w 是对偶问题(DLP)的最优解.

3.3 对偶单纯形法

线性规划**对偶单纯形法**的思想,是对对偶问题用单纯形法.从形式上看,相当于在每步迭代中,原空间中的 x 迭代点要求是基本解(满足 $Ax = b$),并且

$$A^T A_B^{-T} c_B \leqslant c,$$

$$(c - A^T A_B^{-T} c_B)_i \cdot x_i = 0, i = 1, 2, \cdots, n.$$

逐步迭代到 $x \geqslant 0$,从而在原空间得到最优极点.

定义 1 设 A_B 是非奇异的,定义 $w = A_B^{-T} c_B$,如果它满足 $A^T w \leqslant c$,则称 w 为**对偶基本可行解**.

容易验证,对基本解 x(其中 $x_B = A_B^{-1}b, x_N = 0$) 而言,

$$(c - A^{\mathrm{T}}A_B^{-\mathrm{T}}c_B)^{\mathrm{T}}x = c^{\mathrm{T}}x - w^{\mathrm{T}}Ax$$
$$= c_B^{\mathrm{T}}A_B^{-1}b - c_B^{\mathrm{T}}A_B^{-1}b = 0,$$

因此,对偶可行性和补偿松弛条件在这种情况下得以满足.显然,只有 $x_B = A_B^{-1}b \geqslant 0$,原可行性才能得到满足.

对偶单纯形法即每次迭代到对偶基本可行解,直到对应的基本解为基本可行解为止.

对偶单纯形法的步骤如下:

步0(初始步) 求解初始的对偶可行基 B^0、初始对偶基本可行解,并计算相对成本向量

$$w_0 = A_{B^0}^{-\mathrm{T}}c_{B^0}, \quad r_0 = c - A^{\mathrm{T}}A_B^{-\mathrm{T}}c_B,$$

令非基指标集 $N^0 = \{1,2,\cdots,n\} \setminus B^0$,置 $k = 0$.

步1(最优性检验) 计算原空间的基本解

$$(x_k)_{B^k} = A_B^{-1}b, x_{N^k} = 0,$$

若 $(x_k)_{B^k} \geqslant 0$,则停止,x_k 即为最优极点;否则转步2.

步2(选取离基指标) 选取(按布兰德原则)$i^* \in \{1,2,\cdots,m\}$,满足

$$(x_k)_{B_{i^*}^k} < 0.$$

步3 计算 v_k:

$$(v_k)_{N^k} = A_{N^k}^{\mathrm{T}}A_{B^k}^{-\mathrm{T}}\bar{e}_{i^*}, (v_k)_{B^k} = \bar{e}_{i^*}.$$

若 $(v_k)_{N^k} \geqslant 0$,则停止迭代,原问题是不可行的;否则,转步4(其中 $\bar{e}_{i^*} \in \mathbf{R}^m$ 是第 i^* 个单位矢量).

步4(选取进基指标) (按布兰德原则)选取 $j^* \in \{1,2,\cdots,n-m\}$,满足

$$-(r_k)_{N_{j^*}^k}/(v_k)_{N_{j^*}^k} = \min\left\{\frac{-(r_k)_{N_j^k}}{(v_k)_{v_j^k}} \middle| (v_k)_{N_j^k} < 0\right\} = t.$$

步5 计算 r_{k+1}

$$(r_{k+1})_N^k = (r_k)_N^k + t(v_k)_N^k,$$
$$(r_{k+1})_B^k = t\bar{e}_{i^*}.$$

步6(修正原空间基本解) 计算 d_k

$$(d_k)_{B^k} = -A_B^{-1}A_{N_{j^*}},$$

$$(d_k)_{N^k} = \widehat{e}_{j^*},$$

$$x_{k+1} = x_k + \frac{(x_k)_{B_{i^*}^k}}{(v_k)_{N_{j^*}^k}}d_k.$$

置 $B^{k+1} = B^k \cup \{N_{j^*}^k\} \setminus \{B_{i^*}^k\}, N^{k+1} = N^k \setminus \{N_{j^*}^k\} \cup \{B_{i^*}^k\}$.

步7 置 $k = k+1$,转步1.

例2 求解线性规划问题

$$\begin{aligned}
\min \quad & -2x_1 - x_2,\\
\text{s.t.} \quad & x_1 + x_2 + x_3 = 2,\\
& x_1 + x_4 = 1,\\
& x_i \geqslant 0, i = 1,2,\cdots,4.
\end{aligned}$$

解:迭代 0.

步 0　选择 $B^0 = \{1,4\}$, $\boldsymbol{w}^0 = \boldsymbol{A}_{B^0}^{-\mathrm{T}}\boldsymbol{c}_{B^0} = (-2,0)^{\mathrm{T}}$, $\boldsymbol{r}_0 = \boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}_{B^0}^{-\mathrm{T}}\boldsymbol{C}_{B^0} = (0,1,2,0)^{\mathrm{T}}$,可见 $\boldsymbol{w}^0$ 对偶可行($N^0 = \{2,3\}$).

步 1　检查最优性　$(\boldsymbol{x}_0)_{B^0} = \boldsymbol{A}_{B^0}^{-1}\boldsymbol{b} = (2,-1)^{\mathrm{T}}$,相应的原始向量是不可行的.

步 2　选离基指标.由于 $(\boldsymbol{x}_0)_{B_2^0} < 0$,选取 $i^* = 2, B_{i^*}^0 = 4$.

步 3　计算 $\boldsymbol{v}_0$

$$(\boldsymbol{v}_0)_{N^0} = \boldsymbol{A}_{N^0}^{\mathrm{T}}\boldsymbol{A}_{B^0}^{-\mathrm{T}}\bar{\boldsymbol{e}}_2 = (-1,-1)^{\mathrm{T}}, (\boldsymbol{v}_0)_{B^0} = (0,1)^{\mathrm{T}}.$$

步 4　选进基指标

$$t = \min\{-(r_0)_{N_j^0}/(v_0)_{N_j^0} \mid (v_0)_{N_j^0} < 0\} = 1 = -\frac{(r_0)_{N_1^0}}{(v_0)_{N_j^0}},$$

$$j^* = 1, N_{j^*}^0 = 2.$$

步 5　计算 $\boldsymbol{r}_1$

$$\begin{aligned}
(\boldsymbol{r}_1)_{N^0} &= (\boldsymbol{r}_0)_{N^0} + t(\boldsymbol{v}_0)_{N^0}\\
&= (1,2)^{\mathrm{T}} + 1 \times (-1,-1)^{\mathrm{T}} = (0,1)^{\mathrm{T}},\\
(\boldsymbol{r}_1)_{B^0} &= t\bar{\boldsymbol{e}}_{i^*} = 1 \times \bar{\boldsymbol{e}}_2 = (0,1)^{\mathrm{T}}.
\end{aligned}$$

步 6　计算 $\boldsymbol{d}_0$

$$(\boldsymbol{d}_0)_{B^0} = -\boldsymbol{A}_{B^0}^{-\mathrm{T}}\boldsymbol{A}_{N_1^0} = (-1,1)^{\mathrm{T}},$$

$$(\boldsymbol{d}_0)_{N^0} = \hat{\boldsymbol{e}}_1 = (1,0)^{\mathrm{T}},$$

$$\begin{aligned}
\boldsymbol{x}_1 &= \boldsymbol{x}_0 + \frac{(\boldsymbol{x}_0)_{B_2^0}}{(\boldsymbol{v}_0)_{N_1^0}}\boldsymbol{d}_0\\
&= (2,0,0,-1)^{\mathrm{T}} + \frac{-1}{-1} \times (-1,1,0,1)\\
&= (1,1,0,1)^{\mathrm{T}}.
\end{aligned}$$

由于 $\boldsymbol{x}_1 \geqslant 0$,因此 $\boldsymbol{x}_1 = (1,1,0,0)^{\mathrm{T}}$ 即为最优极点.

4　灵敏度分析

标准型线性规划问题(LP)完全由数据$(\boldsymbol{A},\boldsymbol{c},\boldsymbol{b})$来确定.如果$(\boldsymbol{A},\boldsymbol{c},\boldsymbol{b})$在一定范围内发生变化,最优解的变化情况则是人们关心的问题.这就是**灵敏度分析**要解

决或处理的问题.下面仅对三种情况的灵敏度进行分析,即分析 c 发生变化,b 发生变化以及 A 发生变化这三种情形.

4.1 目标函数的改变

说 x^* 是标准型线性规划(LP)的最优解,对应的基指标集与非基指标集分别为 B 和 N.

设 $\bar{c} = c + \alpha c'$,考虑扰动后的线性规划问题

$$(\mathrm{LP}_1)\qquad \begin{cases}\min & \bar{c}^{\mathrm{T}} x,\\ \text{s.t.} & Ax = b,\\ & x \geqslant 0.\end{cases}$$

讨论 α 的范围,使在这一范围之内,x^* 仍是(LP_1)的最优解.

因为可行域没有发生变化,如果 x^* 依然是(LP_1)的最优解,则

$$r_N^{\mathrm{T}} = \bar{c}_N^{\mathrm{T}} - \bar{c}_B^{\mathrm{T}} A_B^{-1} A_N \geqslant 0$$

成立.换言之,有

$$(c_N + \alpha c'_N)^{\mathrm{T}} - (c_B + \alpha c'_B)^{\mathrm{T}} A_B^{-1} A_N \geqslant 0,$$

即 α 满足

$$\alpha r'_N \geqslant - r_N,$$

其中

$$r'_N = c'_N - A_N^{\mathrm{T}} A_B^{-\mathrm{T}} c'_B.$$

定义

$$\underline{\alpha} = \max\left\{\max\left\{\frac{-r_{N_j}}{r'_{N_j}} \mid r'_{N_j} > 0, j = 1,2,\cdots,n-m\right\} - \infty\right\},$$

$$\bar{\alpha} = \min\left\{\min\left\{\frac{-r_{N_j}}{r'_{N_j}} \mid r'_{N_j} < 0, j = 1,2,\cdots,n-m\right\} + \infty\right\}.$$

结论 $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ 时,x^* 是(LP_1)的最优解;最优值 $z^*(\alpha)$ 在此范围内是 α 的线性函数.

可以进一步证明 $z^*(\alpha)$ 是 α 的凹的分段线性函数.

4.2 右端向量的改变

设 x^* 是(LP)的最优解,对应的基指标集与非基指标集分别为 B 和 N.考虑当 b 变化为 $b + \alpha b'$ 时的线性规划问题

$$(\mathrm{LP}_2)\qquad \begin{aligned}&\min \quad c^{\mathrm{T}} x,\\ &\text{s.t.} \quad Ax = b + \alpha b', x \geqslant 0.\end{aligned}$$

讨论 α 的范围,使得在这一范围内(LP_2)的最优基仍是 B.

要保证 B 仍是最优基,α 必须满足

$$\bar{x} = A_B^{-1}(b + \alpha b') \geqslant 0,$$

因为相对成本向量 $\bar{r} = c - A^{\mathrm{T}}A_B^{-\mathrm{T}}c_B \geqslant 0$ 是显然的.

令 $y_B = A_B^{-1}b'$,

定义

$$\underline{\alpha} = \max\left\{\max\left\{\frac{-x_{B_i}^*}{y_{B_i}} \mid y_{B_i} > 0, i = 1,2,\cdots,m\right\}, -\infty\right\},$$

$$\bar{\alpha} = \min\left\{\min\left\{\frac{-x_{B_i}^*}{y_{B_i}} \mid y_{B_i} < 0, i = 1,2,\cdots,m\right\}, +\infty\right\}.$$

结论 当 $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ 时,B 仍是(LP_2)的最优基,此时最优解与最优值均是 α 的线性函数.

4.3 约束矩阵的改变

约束矩阵的改变,其灵敏度分析不是简单的工作,这一节仅考虑三种较简单的情况:增加一个变量,减少一个变量,增加一个约束.像前几节一样,仍然设 x^* 是(LP)问题的最优解,对应的最优基指标集与非基指标集分别是 B 和 N.

情况 1 增加一个变量.设增加变量为 x_{n+1},增加后的线性规划为

$$(\mathrm{LP}_3)\qquad \begin{aligned} &\min \quad c^{\mathrm{T}}x + c_{n+1}x_{n+1},\\ &\text{s.t.} \quad Ax + A_{n+1}x_{n+1} = b, x \geqslant 0, x_{n+1} \geqslant 0,\end{aligned}$$

其中 $c_{n+1} \in \mathbf{R}, A_{n+1} \in \mathbf{R}^m$.

注意到 $(x^*, 0)^{\mathrm{T}}$ 是(LP_3)的基本可行解,若 $r_{n+1} = c_{n+1} - c_B^{\mathrm{T}}A_B^{-1}A_{n+1} \geqslant 0$,则它还是($\mathrm{LP}_3$)的最优解;否则,若 $r_{n+1} < 0$,则必须以 $(x^*, 0)^{\mathrm{T}}$ 为初始基本可行解,继续用单纯形法求解(LP_3).显然,下一迭代 $n+1$ 将进入基指标集中.

情况 2 减少一个变量.设从(LP)中删去 x_k.若 $x_k^* = 0$,则 $x_{I\setminus\{k\}}^*$ 即为新的线性规划的最优解,其中 $I = \{1,2,\cdots,n\}$.若 $x_k^* > 0$,必须算出新的解.一般采用求解下面的第一阶段问题

$$(\mathrm{LP}_4)\qquad \begin{aligned} &\min \quad x_k,\\ &\text{s.t.} \quad Ax = b, x \geqslant 0.\end{aligned}$$

将 x_k 从基变量中删去,显然 x^* 是(LP_4)的一个基本可行解,可用单纯形法求解(LP_4).设最优解为 $\hat{x}$.若最优值 $\hat{x}_k \neq 0$,则必有删掉 x_k 后得到的新的线性规划是不可行的;若 $\hat{x}_k = 0$,则以 $\hat{x}_{I\setminus\{k\}}$ 为初始基本可行解求解新的线性规划问题.

情况 3 增加一个约束.设增加的约束为下述不等式形式:

$$a_{m+1}^{\mathrm{T}}x \leqslant b_{m+1},$$

其中 $a_{m+1} \in \mathbf{R}^n, b_{m+1} \in \mathbf{R}^1$.新的线性规划问题为

$$(\mathrm{LP}_5)\qquad \begin{aligned} &\min \quad c^{\mathrm{T}}x,\\ &\text{s.t.} \quad Ax = b, \quad a_{m+1}^{\mathrm{T}}x \leqslant b_{m+1}, x \geqslant 0.\end{aligned}$$

增加松弛变量 x_{n+1},考虑(LP_5)的等价形式

$$(\mathrm{ALP}_5)\quad \begin{aligned} &\min \quad c^{\mathrm{T}}x, \\ &\mathrm{s.t.} \quad \overline{A}x_{I\cup\{n+1\}} = b', x_{I\cup\{n+1\}} \geqslant 0. \end{aligned}$$

其中

$$\overline{A} = \begin{bmatrix} A & 0 \\ a_{m+1}^{\mathrm{T}} & 1 \end{bmatrix}, \quad b' = \begin{bmatrix} b \\ b_{m+1} \end{bmatrix}, \quad I = \{1,2,\cdots,n\}. \tag{4-1}$$

令 $\overline{B} = B \cup \{n+1\}$,则 $\overline{A}_{\overline{B}}$ 是非奇异的,它的逆为

$$\overline{A}_{\overline{B}}^{-1} = \begin{bmatrix} A_B^{-1} & 0 \\ -(a_{m+1,B}^{\mathrm{T}})A_B^{-1} & 1 \end{bmatrix}, \tag{4-2}$$

因此由

$$\overline{x}'_{\overline{B}} = \overline{A}_{\overline{B}}^{-1}b', \overline{x}'_N = 0 \tag{4-3}$$

定义的 $\overline{x}'$ 是一基本解(不一定可行).

结论 若由(4-3)式定义的 $\overline{x}'$ 是非负的,则它即是(ALP_5)的最优解.

上述结论可证明如下.

检查(ALP_5)的基本可行解 $\overline{x}'$ 关于 $\overline{B}$ 的相对成本向量

$$\overline{r}^{\mathrm{T}} = [c^{\mathrm{T}},0] - [c_B^{\mathrm{T}},0]\overline{A}_{\overline{B}}^{-1}\overline{A},$$

由于 $\overline{r}_{\overline{B}}^{\mathrm{T}} = 0^{\mathrm{T}}$ 及

$$\begin{aligned} \overline{r}_N^{\mathrm{T}} &= c_N^{\mathrm{T}} - (c_B^{\mathrm{T}},0)\begin{bmatrix} A_B^{-1} & 0 \\ -(a_{m+1,B}^{\mathrm{T}})A_B^{-1} & 1 \end{bmatrix}\begin{bmatrix} A_N \\ a_{m+1,N}^{\mathrm{T}} \end{bmatrix} \\ &= c_N^{\mathrm{T}} - c_B^{\mathrm{T}}A_B^{-1}A_N \geqslant 0, \end{aligned}$$

有 $\overline{r}' \geqslant 0$,因此 $\overline{x}'$ 是(ALP_5)的最优解,$\overline{x}'_I$ 即为(LP_5)的最优解.

若 $\overline{x}'_{\overline{B}} \ngeqslant 0$,这种情况可用对偶单纯形法来求解($\mathrm{LP}_5$),因为 $w' = (c_B^{\mathrm{T}}A_B^{-1},0)^{\mathrm{T}}$ 是一对偶基本可行解.

5 卡马卡算法

5.1 卡马卡标准型问题

1984年,卡马卡(Karmarkar)提出求解线性规划的一个内点算法,它是多项式算法.卡马卡考虑下述标准型问题

$$(\mathrm{KLP})\quad \begin{aligned} &\min \quad c^{\mathrm{T}}x, \\ &\mathrm{s.t.} \quad Ax = 0, e^{\mathrm{T}}x = n, x \geqslant 0. \end{aligned}$$

其中 $c \in \mathbf{R}^n, A \in \mathbf{R}^{m,n}, e \in \mathbf{R}^n, e = (1,2,\cdots,1)^{\mathrm{T}}$.卡马卡标准型问题须满足

1° 最优值是0;

2° $e \in \Omega = \{x \in \mathbf{R}^n \mid Ax = 0, e^{\mathrm{T}}x = n, x \geqslant 0\}$;

3° $\begin{bmatrix} A \\ e^{\mathrm{T}} \end{bmatrix}$ 是行满秩矩阵.

定义

$$\Omega_{+} = \{x \in \mathbf{R}^n \mid Ax = 0, e^{\mathrm{T}}x = n, x > 0\},$$

$$\Pi = \{x \geqslant 0 \mid e^{\mathrm{T}}x = n\}.$$

Π 是一单纯形,它的中心是 e,内切球半径 $r_1 = \sqrt{n/(n-1)}$,外接球半径 $r_2 = \sqrt{n(n-1)}$.

5.2　卡马卡算法

先阐述卡马卡算法的思想,再给出算法的迭代步骤.

定义势函数

$$f(x,c) = n\log c^{\mathrm{T}}x - \sum_{j=1}^{n}\log x_j.$$

注意到(下面将要证明):如果 $x \in \Omega_{+}, f(x,c) \leqslant f(x,e) - r, r > 0$,则有 $c^{\mathrm{T}}x \leqslant \exp(-\frac{1}{n}r)c^{\mathrm{T}}e$.显然可以通过求解非线性规划问题

(NLP)　　　$\min\{f(x,c) \mid x \in \Omega\}$

的近似解得到问题(KNP)的近似解.

设第 k 步得到 $x^k \in \Omega_{+}$,作变换 $y = T_{x^k}(x)$:

$$y = T_{x^k}(x) = \frac{nD(x^k)^{-1}x}{e^{\mathrm{T}}D(x^k)^{-1}x},$$

它的逆变换是

$$x = T_{x^k}^{-1}(y) = \frac{nD(x^k)y}{e^{\mathrm{T}}D(x^k)y},$$

其中,$D(x^k) = \mathrm{diag}(x_1^k, x_2^k, \cdots, x_n^k)$,(NLP)等价地变换为

(NLP)$_k$　　　$\min\{f(T_{x^k}^{-1}(y), c) \mid y \in \Omega^k\}$,

其中 $\Omega^k = \{y \geqslant 0 \mid AD(x^k)y = 0, e^{\mathrm{T}}y = n\}$.

由 $f(T_{x^k}^{-1}(y), c) = f(y, D(x^k)c) - \prod_{j=1}^{n}\log x_j^k$,

可得:若 $y \in \Omega_{+}^k = \{y > 0 \mid AD(x^k)y = 0, e^{\mathrm{T}}y = n\}$ 满足

$$f(y, D(x^k)c) \leqslant f(e, D(x^k)c) - \gamma,$$

则 $f(T_{x^k}^{-1}(y), c) - f(x^k, c) \leqslant -\gamma$.可见若每次迭代 k,均可求得 $y^k \in \Omega_{+}^k$ 满足 $f(x^k, D(x^k)c) - f(e, D(x^k)c) \leqslant -\delta$,其中 $\delta > 0$ 是常数.令 $x^{k+1} = T_{x^k}^{-1}(y^k)$,则 $f(x^{k+1}, c) - f(x^k, c) \leqslant -\delta$.取 $x^1 = e$,就可得

$$f(x^{k+1}, c) - f(x^k, c) \leqslant -k\delta.$$

因此 $c^{\mathrm{T}}x^{k+1} \leqslant \exp(-\frac{k}{n}\delta)c^{\mathrm{T}}e$,这正是卡马卡算法多项式复杂性的关键.

关键的问题在于如何求$(\mathrm{NLP})_k$的一个近似解y^k,使它满足上述条件.卡马卡选取从$e\in\Omega_+^k$出发,沿罗森(J.B.Roson)投影方向迭代到Ω_+^k的一个位于Π之内切球内部的一点作为y^k,即

$$y^k = e + \alpha d^k \Big/ \| d^k \|,$$

$$d^k = -P_{B_k}\nabla f(e,c^k),\alpha\in(0,1),$$

其中$c^k = D(x^k)c$,$B_k = \begin{bmatrix} AD(x^k) \\ e^{\mathrm{T}} \end{bmatrix}$,矩阵

$$P_{B_k} = I - B_k^{\mathrm{T}}(B_kB_k^{\mathrm{T}})^{-1}B_k$$

是沿$\mathrm{Range}(B_k^{\mathrm{T}})$到$\mathrm{Null}(B_k)$的直交投影矩阵.

卡马卡算法的步骤如下:

步1 置$x^l = e$,$k=1$,给定精度$\epsilon>0$.

步2 若$c^{\mathrm{T}}x^k\leqslant\epsilon$,终止计算,$x^k$满足精度.

步3 计算$A_k = AD(x^k)$,$c^k = D(x^k)c$.

步4 计算$d^k = -P_{B_k}c^k\left(=P_{A_k}c^k-\dfrac{1}{n}e^{\mathrm{T}}c^k\times e\right)$.

步5 计算$y^k = e+\alpha d^k\Big/\|d^k\|$,$\alpha\in\left(0,\dfrac{1}{2}\right)$.

步6 计算 $x^{k+1} = nD(x^k)y^k\Big/e^{\mathrm{T}}D(x^k)y^k$,置$k=k+1$,转步2.

5.3 卡马卡算法的收敛性

这一节证明卡马卡算法的收敛性.

引理1 $y^k\in\Omega_+^k$.

引理2 $f(T_x^{-1}(y),c) = f(y,c^k)-\sum\limits_{j=1}^{n}\log x_j^k$.

引理3 若$|\mu|\leqslant\alpha<1$,则

$$|\log(1+\mu)-\mu|\leqslant\frac{\mu^2}{2(1-\alpha)^2}.$$

引理4 $\left|\sum\limits_{j=1}^{n}\log y_j^k\right|\leqslant\alpha^2/2(1-\alpha)^2$.

引理5 $n\log c^{k^{\mathrm{T}}}y^k-n\log c^{k^{\mathrm{T}}}e\leqslant-\alpha$.

证明 $c^{k^{\mathrm{T}}}y^k = c^{k^{\mathrm{T}}}e-\alpha\|d^k\|$.注意到$0=\min\{c^{k^{\mathrm{T}}}y\mid y\in\Omega_k\}$及$\Omega_k\subset M_k = B(e,r_2)\cap\{z\mid A_kz=0,e^{\mathrm{T}}z=n\}$,其中$B(e,r_2)=\{z\mid\|z-e\|\leqslant r_2\}$,则有问题

$$\min\{c^{k^{\mathrm{T}}}y\mid y\in M_k\}$$

的最优值非正.注意到上述问题的解为 $e + r_2 d^k / \| d^k \|$,有

$$c^{k^{\mathrm{T}}}(e + r_2 d^k / \| d^k \|) \leqslant 0.$$

据此得
$$\| d^k \| \geqslant \frac{c^{k^{\mathrm{T}}} e}{\sqrt{n(n-1)}} > \frac{c^{k^{\mathrm{T}}} e}{n},$$

从而有
$$c^{k^{\mathrm{T}}} y^k < (1 - \frac{\alpha}{n}) c^{k^{\mathrm{T}}} e,$$

$$n \log c^{k^{\mathrm{T}}} y^k - n \log c^{k^{\mathrm{T}}} e < n \log(1 - \frac{\alpha}{n}) \leqslant -\alpha.$$

引理 6 $f(y^k, c^k) - f(e, c^k) \leqslant -\alpha + \alpha^2/2(1-\alpha)^2 = -\delta.$

定理 1 卡马卡算法所得到的点列 $\{x^k\}$ 满足

$$c^{\mathrm{T}} x^{k+1} \leqslant c^{\mathrm{T}} e \exp(-\frac{k}{n}\delta).$$

证明 由引理 2 与引理 6 有

$$f(x^{j+1}, c) - f(x^j, c) = f(y^j, c^j) - f(e, c^j) \leqslant -\delta,$$

$$f(x^{k+1}, c) - f(e, c) = \sum_{j=1}^{k} (f(x^{j+1}, c) - f(x^j, c)) \leqslant -k\delta.$$

由 $\sum_{j=1}^{k} \log x_j^{k+1} = \log(\prod_{j=1}^{n} x_j^{k+1}) \leqslant 0$ 得

$$\begin{aligned} -k\delta &\geqslant f(x^{k+1}, c) - f(e, c) \\ &= n \log c^{\mathrm{T}} x^{k+1} - \sum_{j=1}^{k} \log x_j^{k+1} - n \log c^{\mathrm{T}} e \\ &\geqslant n \log \frac{c^{\mathrm{T}} x^{k+1}}{c^{\mathrm{T}} e}. \end{aligned}$$

于是有
$$c^{\mathrm{T}} x^{k+1} \leqslant \exp(-\frac{k}{n}\delta) c^{\mathrm{T}} e.$$

下面用定理 1 来说明卡马卡算法的多项式复杂性.

设 L 是问题(KLP)的规模.只要 L 足够大,就可使 $\exp(-L)(c^{\mathrm{T}} e) \approx 0$,即对充分小的 ϵ 有 $\exp(-L)(c^{\mathrm{T}} e) < \epsilon$.由定理 1 知,当 $k > \frac{1}{\delta} nL$ 时,该算法就可在 $c^{\mathrm{T}} x^{k+1} < \epsilon$ 的条件下终止.比如 $\alpha = 1/3$ 时,$\delta = 5/24 > \frac{1}{5}$,则当 $k > 5nL$ 时算法就可终止.

定理 2 在卡马卡算法中取 $\epsilon = \exp(-L) c^{\mathrm{T}} e$,$\alpha \in (0, \frac{1}{2})$,则它在 $O(nL)$ 次迭代终止.

6 多目标线性规划问题及其解

6.1 多目标规划

设 $f:\mathbf{R}^n \to \mathbf{R}^p (p \geqslant 2)$ 是 $\mathbf{R}^n$ 到 $\mathbf{R}^p$ 的向量值映射，$\Omega \subset \mathbf{R}^n$ 是 $\mathbf{R}^n$ 中非空子集，是多目标规划问题的决策集(或称可行域)，称为决策空间. 像集 $Y = f(\Omega) = \{y = f(x) \mid x \in \Omega\} \subset \mathbf{R}^p$ 称为**多目标规划问题的目标空间**. 一个多目标规划问题需要确定决策空间、目标空间以及目标空间中元素之间的优劣关系 D. 这样，多目标规划问题可以记为(Ω, Y, D) 或(Ω, f, D). 多目标规划问题又称为多目标最优化问题，可以写成

$$(\text{VMP}) \qquad \begin{cases} \min & f(x), \\ \text{s.t.} & x \in \Omega \subset \mathbf{R}^n. \end{cases} \tag{6-1}$$

6.2 偏 序

设 $Y \subset \mathbf{R}^p$，若 Y 中元素之间的一个二元关系“$\prec$”满足

(1) 自反性 $\forall y \in Y \Rightarrow y \prec y$，

(2) 传递性 $\forall x, y, z \in Y, x \prec y, y \prec z \Rightarrow x \prec z$，

则称集合 Y 为偏予序集，记为$(Y, \prec)$.

(3) 反对称性 $\forall y, z \in Y, y \prec z, z \prec y \Rightarrow y = z$，

把具有反对称性的偏予序集 Y 称为**偏序集**.

(4) 完全性 $\forall y, z \in Y$，总有 $y \prec z$ 或 $z \prec y$.

把具有完全性的偏序集 Y 称为**全序集**$(Y, \prec)$.

设$(Y, \prec)$为全序集，令

$$D(y) = \{z - y \mid y \prec z, y, z \in Y\},$$

则 $y \prec z \Leftrightarrow z \in y + D(y)$，并称 $D(y)$ 为 Y 在 y 处的控制结构. 若 $D(y)$ 与 y 无关，则记 $D(y) = D$，并称 D 为 Y 上的控制结构. 当 $z \in y + D(y)$(即 $y \prec z$)时，则称 z 受 y 控制，即 $y + D(y)$ 中的一切点均比 y 坏，或不比 y 好.

设 $D \subset \mathbf{R}^p$ 是以零点为顶点的凸锥，且 $D \cap (-D) = \{0\}$，即 D 为点锥，由点锥 D 可定义集合 $Y \subset \mathbf{R}^p$ 上元素之间的一个二元关系“$\prec$”：

$$\forall y_1, y_2 \in Y, y_1 \prec y_2 \Leftrightarrow y_2 - y_1 \in D.$$

可以证明$(Y, \prec)$是个偏序集. 令

$$\mathbf{R}_+^p = \{y = [y_1, y_2, \cdots, y_p] \in \mathbf{R}^p \mid y_j \geqslant 0, j = 1, 2, \cdots, p\},$$

$$\mathbf{R}_{++}^p = \{y = [y_1, y_2, \cdots, y_p] \in \mathbf{R}^p \mid y_j > 0, j = 1, 2, \cdots, p\},$$

则称 $\mathbf{R}_+^p$ 为正锥，称 $\mathbf{R}_{++}^p$ 为严格正锥，在 $\mathbf{R}^p$ 上依正锥 $\mathbf{R}_+^p$ 定义的二元关系称为坐标向量序，简称为**坐标序**. 对于坐标序可引入下列记号：$\forall x, y \in \mathbf{R}^p$，

(1) $x = y \Leftrightarrow x_j = y_j, j = 1,2,\cdots,p$;
(2) $x > y \Leftrightarrow x_j > y_j, j = 1,2,\cdots,p$;
(3) $x \geqslant y \Leftrightarrow x_j \geqslant y_j, j = 1,2,\cdots,p, x \neq y$;
(4) $x \geqq y \Leftrightarrow x_j \geqq y_j, j = 1,2,\cdots,p$.

6.3 有效点与有效解

设像集 $Y = f(\Omega) \subset \mathbf{R}^p$ 与其上的二元关系"$<$"构成全序集$(Y, <)$. D是零点为顶点的凸点锥.

定义 1 若 $y_o \in Y$,且不存在 $y \in Y$,使 $y < y_o, y \neq y_o$(即 $y_o \in y + D \setminus \{o\}$),则称 y_o 为Y的**有效点**(或称 y_o 是Y关于D的锥极点). 若 $x_o \in \Omega$,且不存在 $x \in \Omega$,使 $f(x_o) \in f(x) + D \setminus \{o\}$,则称 x_o 为多目标规划(VMP)的**有效解**(或称 x_o 为(VMP)关于控制结构 D 的锥极解).

定义 2 设点锥 D 的内部 $\mathrm{int}D$ 非空,若 $y_o \in Y$,且不存在 $y \in Y$,使 $y \neq y_o$, $y_o - y \in \mathrm{int}D$,则称 y_o 为 Y 的弱有效点. 若 $x_o \in \Omega$,且不存在 $x \in \Omega$,使 $f(x_o) - f(x) \in \mathrm{int}D, f(x) \neq f(x_o)$,则称 x_o 为多目标规划问题(Ω, f, D)的**弱有效解**.

在有限维实线性空间 $\mathbf{R}^p$ 中,可依坐标序定义其上的二元关系. 由此可分别给出坐标序下像集 $Y \subset \mathbf{R}^p$ 的绝对最优点(这里指绝对最小点)、有效点及弱有效点,以及多目标规划的有效解与弱有效解的定义.

设 $y_o \in Y \subset \mathbf{R}^p$,若 $\forall y \in Y$,有 $y \geqq y_o$,则称 y_o 为 Y 的绝对最小点;若 $\forall y \in Y$,使 $y_o \geqslant y$(或 $y_o > y$),则称 y_o 为 Y 的有效点(或弱有效点).

对于多目标规划$(\Omega, f, \mathbf{R}^p_+)$,设 $x_o \in \Omega$,若 $\forall x \in \Omega$,有 $f(x) \geqq f(x_o)$,则称x_o为多目标规划的绝对最优解(这里指最小解);若不存在 $x \in \Omega$,使 $f(x_o) \geqslant f(x)$(或 $f(x_o) > f(x)$),则称 x_o 为多目标规划$(\Omega, f, \mathbf{R}^p_+)$的有效解(或弱有效解).

绝对最优点(解)、有效点(解)和弱有效点(解)统称为帕雷托(Pareto)最优点(解),分别记为 Y_{ab}, Y_{pa} 和 Y_{wp}(Ω_{ab}, Ω_{pa} 和 Ω_{wp}).

定理 1(接触点定理) 设 $y_o \in Y \subset \mathbf{R}^p$,则

1° $y_o \in Y_{pa}$(有效点集合)$\Leftrightarrow Y \cap (y_o + \mathbf{R}^p_+) = \{y_o\}$,即 y_o 为 Y 与闭锥 $y_o + \mathbf{R}^p_+$ 的唯一接触点.

2° $y_o \in Y_{wp}$(弱有效点集合)$\Leftrightarrow Y \cap (y_o + \mathbf{R}^p_{++}) = \varnothing$,即 Y 与开锥 $y_o + \mathbf{R}^p_{++}$ 没有接触点.

3° $Y_{ab} \subset Y_{pa} \subset Y_{wp} \subset \partial Y$,其中 ∂Y 为 Y 的边界点集合.

相类似,对于多目标规划$(\Omega, f, \mathbf{R}^p_+)$,有 $\Omega_{ab} \subset \Omega_{pa} \subset \Omega_{wp} \subset \partial\Omega$.

6.4 多目标线性规划问题

若多目标线性规划(Ω, f, D)中,向量值目标函数与所有约束函数都是决策变量的线性函数,则称该规划问题为多目标线性规划问题,或称为多目标线性最优化

问题.它的标准形式可以写成

$$(\text{VLP})\qquad \begin{array}{ll} \min & f(\boldsymbol{x}) = \boldsymbol{c}\boldsymbol{x}, \\ \text{s.t.} & A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geqslant \boldsymbol{0}. \end{array} \tag{6-2}$$

其中,$\boldsymbol{c} = (c_{ij})_{p\times n} \in \mathbf{R}^{p\cdot n}$、$A = (a_{ij})_{m\times n} \in \mathbf{R}^{m\cdot n}$,$\boldsymbol{b} \in \mathbf{R}^m$、$x \in \mathbf{R}^n$.多目标线性规划(VLP)的决策空间(或称为(VLP)的可行域)为

$$\Omega = \{\boldsymbol{x} \in \mathbf{R}^n \mid A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geqslant \boldsymbol{0}\}.$$

可见,多目标线性规划(VLP)的可行域与标准单目标线性规划(LP)(见(1-1)式)的可行域完全相同,因此单目标线性规划(LP)及其可行域 Ω 有关的所有基本概念都可移植到多目标线性规划(VLP)中来.例如,多目标线性规划(VLP)中的基本变量、非基本变量、基本可行解、可行域 Ω 的极点(顶点)、极方向等,都与单目标线性规划(LP)中的相同.只是多目标线性规划问题的解与单目标线性规划(LP)的解有所不同.

定义 3　在多目标线性规划(VLP)中,若 $\boldsymbol{x}_o \in \Omega_{pa}$,且 $\boldsymbol{x}_o$ 为(VLP)的基本可行解,则称 $\boldsymbol{x}_o$ 为多目标线性规划问题(VLP)的非劣极点.若 $\boldsymbol{x}^1$、$\boldsymbol{x}^2 \in \Omega_{pa}$ 为可行域 Ω 的相邻基本可行解,$\forall \alpha \in [0,1]$,$\boldsymbol{x} = \alpha\boldsymbol{x}^1 + (1-\alpha)\boldsymbol{x}^2 \in \Omega_{pa}$,则称 $\boldsymbol{x}^1,\boldsymbol{x}^2 \in \Omega$ 是(VLP)的相邻非劣极点,$\boldsymbol{x}^1$ 与 $\boldsymbol{x}^2$ 的连线称为(VLP)的非劣棱.若 Ω 边界面 F 中每一点都是多目标线性规划问题(VLP)的非劣点,即(VLP)的有效解,则称 F 是(VLP)的非劣面;若 F 不包含在其他非劣面中,则 F 是(VLP)的一个最大非劣面.

例 1　设有如下多目标线性规划问题$(\text{VLP})_1$:

$$\min\quad f(\boldsymbol{x}) = \boldsymbol{c}\boldsymbol{x} = \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$$\text{s.t.}\quad A\boldsymbol{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leqslant \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\boldsymbol{x} = (x_1, x_2) \geqslant 0,$$

则问题$(\text{VLP})_1$的可行域(或称决策空间)为

$$\Omega_1 = \{\boldsymbol{x} = (x_1, x_2) \in \mathbf{R}^2 \mid 0 \leqslant x_1 \leqslant 1, 0 \leqslant x_2 \leqslant 1\},$$

显然,Ω_1 的极点为 $O = (0,0)$,$A = (1,0)$,$B = (1,1)$,$D = (0,1)$.这样多目标规划问题$(\text{VLP})_1$的像集为

$$Y_1 = f(\Omega_1) = \{\boldsymbol{y} = (y_1, y_2)^{\mathrm{T}} = \boldsymbol{c}\boldsymbol{x} \mid \boldsymbol{x} \in \Omega_1\}.$$

像集 Y_1 的顶点为

$$O' = f(0) = (0,0), A' = f(A) = (1,-1),$$
$$B' = f(B) = (3,-2), D' = f(D) = (2,-1).$$

由接触点定理可知,$(\text{VLP})_1$ 的有效解集 Y_{pa_1} 与弱有效解集 Y_{wpa_1} 为

$$Y_{pa_1} = Y_{wpa_1} = \text{折线 } O'A'B'.$$

因 $f(\boldsymbol{x}) = \boldsymbol{c}\boldsymbol{x}$ 是线性变换,所以折线 $O'A'B'$ 的原像为折线 OAB.故(VLP)的有效解集 Ω_{pa_1}、弱有效解集 Ω_{wpa_1} 为

$$\Omega_{pa_1} = \Omega_{wpa_1} = \text{折线 } OAB.$$

7　多目标线性规划问题的解法

多目标规划问题的解法分为间接解法与直接解法两类. 间接解法又可分为转化为单目标优化方法、转化为多个单目标优化方法及非统一模型方法等. 直接解法是指不转化为单目标规划问题,直接求解多目标规划问题的一类方法. 到目前为止,直接解法的研究成果比较少,且只对单变量、可行域为有限域或多目标线性规划等多目标规划问题进行研究.

7.1　间 接 解 法

设多目标规划问题(VMP)由(6-1)式给出,其决策空间为 $\Omega \subset \mathbf{R}^n$,像集 $Y = f(\Omega) \subset \mathbf{R}^p (p \geqslant 2)$ 为目标空间. 间接解法主要有以下几种.

1. 主要目标法

依实际情况,从多个目标 $f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_p(\boldsymbol{x})$ 中选出一个主要目标,比如选 $f_1(\boldsymbol{x})$ 为主要目标. 对其余的 $p-1$ 个目标给出一组允许的限值 $a_j, j = 2,3,\cdots,p$. 这样把多目标规划问题(6-1)就可转化为单目标规划问题:

$$
\text{(SP)}\qquad \begin{aligned} &\min \quad f_1(\boldsymbol{x}), \\ &\text{s.t.} \quad f_j(\boldsymbol{x}) \leqslant a_j, \quad j = 2,3,\cdots,p, \\ &\qquad\quad \boldsymbol{x} \in \Omega \subset \mathbf{R}^n. \end{aligned}
$$

定理 1　若 $\boldsymbol{x}_o \in \Omega$ 为单目标规划问题(SP)的最优解,则 $\boldsymbol{x}_o \in \boldsymbol{\Omega}_{wp}$,即 $\boldsymbol{x}_o$ 为多目标规划问题(VMP)的弱有效解.

2. 评价函数法

对多目标规划问题(VMP)引入评价函数 $u: Y \subset \mathbf{R}^p \to \mathbf{R}^1$ 把多目标规划问题(VMP)转化为单目标规划问题

$$
\text{(SP)}\qquad \begin{aligned} &\min \quad F(\boldsymbol{x}) = u(f(\boldsymbol{x})), \\ &\text{s.t.} \qquad \boldsymbol{x} \in \Omega. \end{aligned}
$$

定理 2　设评价函数 $u(f(\boldsymbol{x}))$ 是定义在像集 $Y = f(\Omega) \subset \mathbf{R}^p$ 上的函数,若 $u(f(\boldsymbol{x}))$ 是 $f(\boldsymbol{x}) \in Y$ 的严格单增函数(或单增函数),则单目标规划问题(SP)的最优解 $\boldsymbol{x}_o$ 为多目标规划问题(VMP)的有效解(或为弱有效解).

设有权向量集合

$$
W_p = \{\boldsymbol{w} \in \mathbf{R}^p \mid \boldsymbol{w} \geqslant 0, \sum_{i=1}^{p} w_i = 1\},
$$

$$
W_p^o = \{\boldsymbol{w} \in \mathbf{R}^p \mid \boldsymbol{w} > 0, \sum_{i=1}^{p} w_i = 1\},
$$

适当选取权系数 $\boldsymbol{\lambda} \in W_p$(或 $\boldsymbol{\lambda} \in W_p^o$),可构成如下的评价函数:

$$
F(\boldsymbol{x}) = u(f(\boldsymbol{x})) = \boldsymbol{\lambda}^{\mathrm{T}} f(\boldsymbol{x}).
$$

这样，多目标规划问题(VMP)可转化为单目标规划问题

$$(\mathrm{SP})_\lambda \qquad \begin{aligned} &\min \quad F(\boldsymbol{x}) = u(f(\boldsymbol{x})) = \sum_{j=1}^{p} \lambda_j f_j(\boldsymbol{x}), \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega \subset \mathbf{R}^n. \end{aligned}$$

若 $\boldsymbol{\lambda} \in W_p^o$，则 $u(f(\boldsymbol{x}))$ 为 $f(\boldsymbol{x})$ 的严格单增函数，由定理1可知，若 $\boldsymbol{x}_o$ 为单目标规划问题$(\mathrm{SP})_\lambda$的最优解，则 $\boldsymbol{x}_o \in \Omega_{pa}$，即 $\boldsymbol{x}_o$ 为多目标规划问题(VMP)的有效解．与此类似，若 $\boldsymbol{\lambda} \in W_p$，则单目标规划问题$(\mathrm{SP})_\lambda$的最优解 $\boldsymbol{x}_o$ 为多目标规划问题(VMP)的弱有效解．

如果多目标规划问题(VMP)的评价函数定义为

$$F(\boldsymbol{x}) = u(f(\boldsymbol{x})) = \max\{f_j(\boldsymbol{x}) \mid j = 1,2,\cdots,p\},$$

或适当选取权系数 $\boldsymbol{\lambda} \in W^o$，令

$$F(\boldsymbol{x}) = u(f(\boldsymbol{x})) = \max\{\lambda_j f_j(\boldsymbol{x}) \mid j = 1,2,\cdots,p\},$$

则多目标规划问题(VMP)可化为单目标规划问题

$$(\mathrm{P}_\lambda) \qquad \begin{aligned} &\min \quad F(\boldsymbol{x}) = \max\{\lambda_j f_j(\boldsymbol{x}) \mid j = 1,2,\cdots,p\}, \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega \subset \mathbf{R}^n. \end{aligned}$$

由定理2可以证明：$\forall \lambda \in W_p^o$，单目标规划问题$(\mathrm{P}_\lambda)$的最优解 $\boldsymbol{x}_o$ 必为多目标规划问题(VMP)的有效解．

如果对每个目标函数 $f_j(\boldsymbol{x})$ 都能选出目标值 f_j^o，使

$$f_j^o \leqslant \min\{f_j(\boldsymbol{x}) \mid \boldsymbol{x} \in \Omega\}, \quad j = 1,2,\cdots,p,$$

则称 $f^o = (f_1^o, f_2^o, \cdots, f_p^o)$ 为多目标规划问题(VMP)的理想点．这样，应用 $\mathbf{R}^p$ 中的范数 $\|\cdot\|_2$，可构造多目标规划问题(VMP)的一个评价函数，从而转化为单目标问题

$$(\mathrm{P}_2) \qquad \begin{aligned} &\min \quad F(\boldsymbol{x}) = u(f(\boldsymbol{x})) = \| f(\boldsymbol{x}) - f^o \|_2, \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega \subset \mathbf{R}^n. \end{aligned}$$

依定理2可以证明：单目标规划问题(P_2)的最优解 $\boldsymbol{x}_o$ 为多目标规划问题(VMP)的有效解．

3．分层排序法

依各子目标的重要程度将其排序，设排序结果为 $f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_p(\boldsymbol{x})$．然后依次求解下列单目标规划问题：

$$(\mathrm{P}_1) \qquad \begin{aligned} &\min \quad f_1(\boldsymbol{x}), \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega \subset \mathbf{R}^n, \end{aligned}$$

设单目标规划问题(P_1)的最优值为 f_1^*，令 $\Omega_1 = \{\boldsymbol{x} \in \Omega \mid f_1(\boldsymbol{x}) = f_1^*\}$．

$$(\mathrm{P}_2) \qquad \begin{aligned} &\min \quad f_2(\boldsymbol{x}), \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega_1, \end{aligned}$$

设单目标规划问题(P_2)的最优值为 f_2^*，令 $\Omega_2 = \{\boldsymbol{x} \in \Omega_1 \mid f_2(\boldsymbol{x}) = f_2^*\}$．

依此类推，

$$(\mathrm{P}_p) \qquad \begin{aligned} &\min \quad f_p(\boldsymbol{x}), \\ &\text{s.t.} \quad \boldsymbol{x} \in \Omega_{p-1}, \end{aligned}$$

设问题(P_p)的最优值为f_p^*,令$\Omega_p = \{x \in \Omega_{p-1} \mid f_p(x) = f_p^*\}$.

定理 3 依上述分层排序求解多个单目标规划问题(P_1),(P_2),…,(P_p).若$x_o \in \Omega_p$,则$x_o \in \Omega_{pa}$,即x_o为多目标规划问题(VMP)的有效解.

7.2 多目标线性规划基本定理

设多目标线性规划问题(VLP)(由(6-2)式给出)的可行域Ω有q个极点$x^1,x^2,\cdots,x^q \in \mathbf{R}^n$,令$Q = \{x^1,x^2,\cdots,x^q\}$.把$Q$的凸包及相对内部分别记为

$$H(Q) = \{x \in \mathbf{R}^n \mid x = \sum_{j=1}^{q}\lambda_j x^j, \lambda \in W_q\},$$

$$H^o(Q) = \{x \in \mathbf{R}^n \mid x = \sum_{j=1}^{q}\lambda_j x^j, \lambda \in W_q^o\}.$$

依单目标线性规划(LP)的基本定理和多目标规划问题(VMP)的有效解概念可以证明定理 4.

定理 4 设$x^1,x^2 \in H(Q)$,若$x^1 \bar{\in} \Omega_{pa}$,$(x^1,x^2) \cap \Omega_{pa} = \varnothing$,其中,$(x^1,x^2) = \{x \mid x = \alpha x^1 + (1-\alpha)x^2, \forall \alpha \in (0,1)\}$,若$x^o \in H(Q)$,且$x^o \bar{\in} \Omega_{pa}$,则$H^o(Q) \cap \Omega_{pa} = \varnothing$.

由定理 4 易证明定理 5.

定理 5 若$x^o \in H^o(Q)$是多目标线性规划问题(VLP)的非劣解,则$\forall x \in H(Q)$,x均为多目标线性规划问题(VLP)的非劣解.

引入权系数$\lambda \in W_p$或$\lambda \in W_p^o$,可把多目标线性规划问题(VLP)转化为如下的单目标线性规划问题:

$$\mathrm{P}(\lambda) \qquad \begin{aligned} \min\ & \lambda^{\mathrm{T}} f(x) = \lambda^{\mathrm{T}} c x, \\ \text{s.t.}\ & Ax = b, x \geqslant 0. \end{aligned}$$

设多目标线性规划问题(VLP)的可行域Ω有r个极点$x^1,x^2,\cdots,x^r$和$q-r$个极方向$x^{r+1},x^{r+2},\cdots,x^q$,由这些极点与极方向组成的集合记为$Q$,把$Q$生成的凸多面体集和其相对内部记为

$$P(Q) = \{x \in \mathbf{R}^n \mid x = \sum_{j=1}^{r}\lambda_j x^j + \sum_{j=r+1}^{q}\mu_j x^j, \lambda \in W_r, \mu_j \geqslant 0\},$$

$$P^o(Q) = \{x \in \mathbf{R}^n \mid x = \sum_{j=1}^{r}\lambda_j x^j + \sum_{j=r+1}^{q}\mu_j x^j, \lambda \in W_q^o, \mu_j > 0\}.$$

定理 6 若$x^o \in P^o(Q)$是多目标线性规划问题(VLP)的非劣解,则$\forall x \in P(Q)$,x都是多目标线性规划问题(VLP)的非劣解,且存在$\lambda^o \in W_p$,使每个$x \in P(Q)$都是单目标线性规划问题$\mathrm{P}(\lambda^o)$的最优解.

设多目标线性规划问题(VLP)的有效解的集合仍记为Ω_{pa}(又称为(VLP)的非劣点集合).

定理 7 在多目标线性规划问题(VLP)中,若存在$\lambda^o \in W_p^o$,使$(\lambda^o)^{\mathrm{T}} cx$ = 常

数,则 $\Omega = \Omega_{pa}$. 否则 $\Omega_{pa} \subset \bigcup_{j=1}^{T} F_j$,其中 F_j 为可行域 Ω 的面,T 是 Ω 中面的个数. Ω 中面 F 是非劣的充要条件:存在 $\lambda^o \in W_p^o$,使每个 $x \in F$ 都是单目标线性规划问题 $P(\lambda^o)$ 的最优解.

设 $E(\Omega_{pa})$ 是多目标线性规划问题(VLP)的非劣极点与非劣极方向的集合,显然 $E(\Omega_{pa}) \subset \Omega_{pa}$. $\forall x^i \in E(\Omega_{pa})$. 定义集合

$$W(x^i) = \{\lambda \in W_p^o \mid x^i \text{ 是问题 } P(\lambda) \text{ 的最优解}\},$$

显然,$\forall x^i \in E(\Omega_{pa}), W(x^i) \neq \varnothing$,且可以证明:

1° $\bigcup_{x^i \in E(\Omega_{pa})} W(x^i) = W^o$.

2° 若面 F 是 $S \subset E(\Omega_{pa})$ 的生成面,则面 F 是非劣的充要条件为 $\bigcap_{x^i \in E(\Omega_{pa})} W(x^i) \neq \varnothing$;而 F 是最大非劣面的充要条件为 $\forall S', S \subset S' \subset E(\Omega_{pa})$,

$$\bigcap_{x^i \in S} W(x^i) \neq \varnothing, \text{且} \bigcap_{x^i \subsetneq S} W(x^i) = \varnothing.$$

7.3 多目标线性规划问题的解法

因多目标线性规划问题(VLP)与单目标线性规划问题(LP)有相同的可行域 $\Omega \subset \mathbf{R}^n$,而 Ω 是凸多面体,Ω 的表面只能由 Ω 的极点和棱所确定的面组成. 由7.2节关于多目标线性规划问题(VLP)的基本定理可知,多目标线性规划问题(VLP)的非劣解集 Ω_{pa} 除非特殊情况是 Ω 外,只能出现在 Ω 的表面上. 若 Ω 的面或棱的内点为(VLP)的非劣解,则整个面或棱也一定是非劣的. 其次,Ω_{pa} 是连通的,所以用单纯形法求多目标线性规划(VLP)非劣解集 Ω_{pa} 的主要步骤如下:

步1 求初始非劣极点. 有两种方法:其一是选 $\lambda^o \in W_p^o$,求问题 $P(\lambda^o)$ 的最优解 x_o,则 x_o 必为(VLP)的非劣极点. 若 $P(\lambda^o)$ 不存在有限最优解,可重新选取 $\lambda^o \in W_p^o$,重复上一过程. 其二是先找出(VLP)的一个基本可行解 x_o,然后用非劣性检验定理来判别 x_o 的非劣性.

步2 相邻极点和棱的搜索. 应用单纯形表及进基与离基原则进行旋转变换,求出相邻基本可行解,并用非劣性检验定理判别其是否为(VLP)的非劣极点.

步3 求最大非劣面. 依定理7和最大非劣面定义直接求出. 求出所有最大非劣面后再求其并集,就得到非劣解集 Ω_{pa}

8 多目标线性规划的对偶性

设有多目标规划问题:

$$(\text{VP}) \quad \begin{aligned} &\min \ f(x), \\ &\text{s.t.} \ \ g(x) \leqslant 0, h(x) = 0, x \geqslant 0. \end{aligned}$$

其中 $x \in \mathbf{R}^n, f:\mathbf{R}^n \to \mathbf{R}^p(p \geqslant 2), g:\mathbf{R}^n \to \mathbf{R}^m, h:\mathbf{R}^n \to \mathbf{R}^s$ 均为可微向量函数. 令 $u \in \mathbf{R}^m, V \in \mathbf{R}^s, H_j = u^{\mathrm{T}}g(x) + v^{\mathrm{T}}h(x), j = 1,2,\cdots,p$,则称函数

$$L(x,u,v) = f(x) + H(x,u,v) - \nabla_x(f(x) + H(x,u,v))^{\mathrm{T}}x$$

为问题(VP)的拉格朗日函数.其中 $H(x,u,v) = [H_1, H_2, \cdots, H_p]$.利用拉格朗日函数建立多目标规划问题(VP)的对偶规划

$$
(\mathrm{VD})\quad
\begin{aligned}
\max\quad & G(y,u,v) = L(y,u,v) \\
& \qquad = f(y) + H(y,u,v) - \nabla_y(f(y) + H(y,u,v))^{\mathrm{T}}y, \\
\text{s.t.}\quad & \lambda^{\mathrm{T}}\nabla f(y) + u^{\mathrm{T}}\nabla g(y) + v^{\mathrm{T}}\nabla h(y) \geqslant 0, \\
& u \geqslant 0(\text{其中 } \lambda \in W_p \text{ 或 } \lambda \in W_p^o).
\end{aligned}
$$

对于多目标线性规划问题(LVP),有

$$
(\mathrm{LVP})\quad
\begin{aligned}
\min\quad & f(x) = cx, \\
\text{s.t.}\quad & Ax \geqslant Bw, x \geqslant 0.
\end{aligned}
$$

其中 $x \in \mathbf{R}^n, A = (a_{ij})_{m\times n} \in \mathbf{R}^{m\times n}, B = (b_{ij})_{m\times p} \in \mathbf{R}^{m\times p}, c = (c_{ij})_{p\times n} \in \mathbf{R}^{p\times n}, W \in \mathbf{R}^p$.在问题(LVP)中,若令 $g(x) = -Ax + BW$,并应用拉格朗日对偶理论可得问题(LVP)的对偶规划为

$$
(\mathrm{LVD})\quad
\begin{aligned}
\max\quad & G(u) = B^{\mathrm{T}}u, \\
\text{s.t.}\quad & u^{\mathrm{T}}A \leqslant \lambda^{\mathrm{T}}c, u \geqslant 0.
\end{aligned}
$$

其中 $u \in \mathbf{R}^m, \lambda \in \mathbf{R}^p$.

定理1(弱对偶定理) 对于线性多目标规划问题(LVP)及其对偶规划(LVD)的任意可行解 x 与 u,总有 $\lambda^{\mathrm{T}}cx \geqslant W^{\mathrm{T}}(B^{\mathrm{T}}u)$ 成立.

定理2(直接对偶定理) 设 $W \in W_p$(或 W_p^o),若 x 是问题(LVP)对应于 W 的弱有效解(或有效解),则存在 $\lambda \in W_p, u \in \mathbf{R}^m$,使 u 是对偶规划(LVD)对应 λ 的弱有效解(或有效解),且有 $\lambda^{\mathrm{T}}cx = W^{\mathrm{T}}B^{\mathrm{T}}u$ 成立.

定理3(逆对偶定理) 设 $\lambda \in W_p$(或 W_p^o),若 u 是对偶规划(LVD)对应 λ 的弱有效解(或有效解),则存在 $W \in W_p, x \in \mathbf{R}^n$,使 x 是原始规划(LVP)对应于 W 的弱有效解(或有效解),且有 $\lambda^{\mathrm{T}}cx = W^{\mathrm{T}}B^{\mathrm{T}}u$ 成立.

参考文献

1 Dantzig G B. Linear programming and extensions. New Jersey: Princeton University Press, 1963.

2 方述诚,普森普拉 S.线性优化及扩展:理论与算法.汪定伟,王梦光译.北京:科学出版社,1994.

3 管梅谷,郑汉鼎.线性规划.济南:山东科学技术出版社,1983.

4 林锉云,董加礼.多目标优化的方法与理论.吉林:吉林教育出版社,1992.

·经济数学卷·

第 7 篇

非线性规划

编　者　王长钰　王宜举
审校者　施光燕

目　　录

引　言

非线性规划是近30年来随着计算机的发展而迅速发展起来的运筹学的一个主要分支.其主要内容是关于最优化问题的理论与算法研究.在理论方面,非线性规划从数学的其他若干分支中汲取营养,逐步形成了自身的学科特色.如非线性规划的最优性判定、鞍点、对偶以及稳定性等理论.在应用方面,非线性规划为系统优化与决策管理提供了强有力的工具,因而在经济管理、生产组织与计划、交通运输、工程技术、科学实验以及军事科学等领域中,都得到了广泛而有效的应用.

本篇主要介绍非线性最优化问题.它的一般模型可以表示为:在若干可微函数构成的等式与不等式约束之下,求一个可微的目标函数的极小值.对于求目标函数极大值的情况,等价于求目标函数相反数的极小值.本篇以介绍方法为主,辅以数值算例,目的在于方便读者掌握应用.

1　无约束最优化问题

1.1　无约束最优化问题的最优性条件

1.1.1　极值点概念

无约束最优化问题可以表示为

$$\min \quad f(\boldsymbol{x}), \boldsymbol{x} \in \mathbf{R}^n, \tag{1-1}$$

其中 $\mathbf{R}^n$ 表示 n 维欧氏空间,$f(\boldsymbol{x})$ 为连续可微函数.

1. 局部极小点

设 $\boldsymbol{x}^* \in \mathbf{R}^n$,若存在 $\delta > 0$,使得对于所有满足 $\boldsymbol{x} \in \mathbf{R}^n$ 和 $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$ 的 $\boldsymbol{x}$ 均有 $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*)$,则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的**局部极小点**.

2. 严格局部极小点

设 $\boldsymbol{x}^* \in \mathbf{R}^n$,若存在 $\delta > 0$,使得对于所有满足 $\boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{x} \neq \boldsymbol{x}^*, \|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$ 的 $\boldsymbol{x}$ 均有 $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$,则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的**严格局部极小点**.

3. 整体极小点

设 $\boldsymbol{x}^* \in \mathbf{R}^n$,若 $\forall \boldsymbol{x} \in \mathbf{R}^n$,有 $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*)$ 成立,则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的**整体极小点**.

4. 严格整体极小点

设 $\boldsymbol{x}^* \in \mathbf{R}^n$,若 $\forall \boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{x} \neq \boldsymbol{x}^*$,有 $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ 成立,则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的**严格整体极小点**.

1.1.2 最优性条件

下面用 $\nabla f(\boldsymbol{x})$ 表示 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 点处的梯度，亦即 $\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right)^{\mathrm{T}}$.

定理 1(一阶必要条件) 设 $\boldsymbol{x}^* \in \mathbf{R}^n$ 是(1-1)式的局部极小点，则 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 点的梯度 $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

用 $\nabla^2 f(\boldsymbol{x})$ 表示 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 点处的黑塞(Hessian)矩阵，即

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{bmatrix}.$$

定理 2(二阶必要条件) 设 $f(\boldsymbol{x})$ 是二阶连续可微函数. 若 $\boldsymbol{x}^* \in \mathbf{R}^n$ 是(1-1)式的局部极小点，则 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处的梯度 $\nabla f(\boldsymbol{x}^*) = 0$，并且 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处的黑塞矩阵 $\nabla^2 f(\boldsymbol{x}^*)$ 半正定.

定理 3(二阶充分条件) 设 $f(\boldsymbol{x})$ 是二阶连续可微函数. 若梯度 $\nabla f(\boldsymbol{x}^*) = 0$，并且黑塞矩阵 $\nabla f(\boldsymbol{x}^*)$ 正定，则 $\boldsymbol{x}^*$ 是(1-1)式的严格局部极小点.

1.1.3 目标函数是凸函数的最优性条件

1. 凸集

设集合 $S \subset \mathbf{R}^n$，如果对于任意的 $x_1, x_2 \in S, \alpha \in (0,1)$，有

$$\alpha x_1 + (1-\alpha) x_2 \in S,$$

则称 S 是凸集.

2. 凸函数

设 $S \subset \mathbf{R}^n$ 是非空凸集，$f(x)$ 是定义在 S 上的函数. 如果对于任意的 $x_1, x_2 \in S, \alpha \in (0,1)$，有

$$f(\alpha x_1 + (1-\alpha) x_2) \leqslant \alpha f(x_1) + (1-\alpha) f(x_2)$$

成立，则称 $f(\boldsymbol{x})$ 是 S 上的凸函数. 如果上述严格不等式成立，则称 $f(\boldsymbol{x})$ 是 S 上的严格凸函数.

定理 4 设 $f(\boldsymbol{x}): \mathbf{R}^n \to \mathbf{R}$ 是连续可微的凸函数，则 $\boldsymbol{x}^*$ 是(1-1)式整体极小点的充分必要条件是

$$\nabla f(\boldsymbol{x}^*) = 0.$$

1.2 最速下降算法

最速下降算法以负梯度方向作为极小化算法的下降方向，故又称为**负梯度法**. 它是无约束最优化中最简单的方法，也是所有可微最优化方法的基础.

算法 1(最速下降算法)

步 1　给出 $x^0 \in \mathbf{R}^n$，允许误差 $\varepsilon > 0, k := 0$.

步 2　计算 $d^k = -\nabla f(x^k)$，若 $\|\nabla f(x^k)\| \leqslant \varepsilon$，则算法终止，$x^k$ 是近似最优解；否则转步 3.

步 3　进行一维精确搜索，即求步长 α_k，满足

$$f(x^k + \alpha_k d^k) = \min\{f(x^k + \alpha d^k) \mid \alpha \geqslant 0\}.$$

步 4　$x^{k+1} := x^k + \alpha_k d^k, k := k + 1$，转步 2.

下面给出算例

例 1　求函数 $f(x) = 1/3x_1^2 + 1/2x_2^2$ 的极小值点.

解　取初始点 $x^0 = (3,2)^{\mathrm{T}}$，直接计算得

$$\nabla f(x^0) = (2,2)^{\mathrm{T}}, d^0 = (-2, -2)^{\mathrm{T}}.$$

求　$f(x^0 + \alpha d^0) = 10/3\alpha^2 - 8\alpha + 5$ 的极小值点得 $\alpha_0 = 6/5$.

$$x^1 = x^0 + \alpha_0 d^0 = (3/5, -2/5)^{\mathrm{T}},$$

$$\nabla f(x^1) = (2/5, -2/5)^{\mathrm{T}}.$$

用同样的迭代步骤进行计算，得到

$$x^2 = (3/5^2, -2/5^2), \nabla f(x^2) = (2/5^2, 2/5^2)$$

一般地，

$$x^k = (3/5^k, (-1)^{k-1}2/5^k),$$

$$\nabla f(x^k) = (2/5^k, (-1)^k 2/5^k).$$

当迭代到一定程度时，即可使$\nabla f(x^k)$ 的模小于等于给定的误差. 事实上，$f(x)$ 的极小值点是$(0,0)^{\mathrm{T}}$.

算法分析　算法 1 具有全局收敛性和线性的收敛速度. 但必须指出，该算法中的下降方向仅对每一个迭代点而言是“最速下降”的，对整个迭代过程总体而言却并非“最速下降”. 特别当目标函数的等值线是一个椭圆（球）时，最速下降法将会出现“锯齿现象”，此时下降就十分缓慢了.

1.3　牛顿算法

1.3.1　牛顿算法

牛顿算法的基本思想是利用目标函数的二次泰勒展开式，并将其极小化，以它作为目标函数极小化的近似.

算法 2（牛顿算法）

步 1　给出 $x^0 \in \mathbf{R}^n$，允许误差 $\varepsilon > 0, k := 0$.

步 2　求$\nabla f(x^k)$，若 $\|\nabla f(x^k)\| \leqslant \varepsilon$，则算法终止，$x^k$ 作为$f(x)$ 的近似最优解；否则转步 3.

步 3　$x^{k+1} := x^k - (\nabla^2 f(x^k))^{-1}\nabla f(x^k), k := k + 1$，转步 2.

例 2　求 $f(x) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$ 的极小值点.

解　取初始点 $x^0 = (0,3)^{\mathrm{T}}$，通过计算得

$$\nabla f(x^0) = (-44,24)^{\mathrm{T}},$$

$$\nabla^2 f(x^0) = \begin{bmatrix} 50 & -4 \\ -4 & 8 \end{bmatrix}, \quad (\nabla^2 f(x^0))^{-1} = \frac{1}{384}\begin{bmatrix} 8 & 4 \\ 4 & 50 \end{bmatrix},$$

$$x^1 = x^0 - (\nabla^2 f(x^0))^{-1} \nabla f(x^0) = (0.67, 0.33)^{\mathrm{T}}.$$

按照同样的步骤，只要$\nabla^2 f(x^k)$非奇异而且$\nabla f(x^k)$不为零，就可以迭代下去，得到点列为 $x^0, x^1, x^2, \cdots$

相应的迭代结果见表 1-1.

表 1-1

k	0	1	2	3	4	5	6
x^k	0 3.0	0.67 0.33	1.11 0.56	1.41 0.70	1.74 0.80	1.74 0.87	1.83 0.91
$f(x^k)$	52	3.13	0.63	0.12	0.02	0.05	0.0009
$\nabla f(x^k)$	−44 24	−9.39 −0.04	−2.84 0.04	0.80 −0.04	0.21 0.56	−0.07 0	0 −0.04

通过观察容易看出，函数 $f(x)$ 在 $x^* = (2,1)^{\mathrm{T}}$ 处达到极小值.

算法分析 当目标函数为正定二次函数时，用算法 2 一次迭代后便可得到极小值点. 对于非二次函数，由于它们在极小值点附近往往和二次函数很近似，因此当初始点靠近极小值点时，算法的收敛速度一般是很快的，此时算法具有局部收敛性与二阶收敛速度. 但当初始点与极小点相距较远时，算法 2 就不能保证有较好的收敛特性，为克服这个缺点，人们提出了阻尼牛顿算法.

1.3.2 阻尼牛顿算法

算法 3(**阻尼牛顿算法**)

步 1 取 $x^0 \in \mathbf{R}^n$，允许误差 $\epsilon > 0$，$k := 0$.

步 2 若 $\| \nabla f(x^k) \| \leqslant \epsilon$，则算法终止，$x^k$ 作为近似极小点；否则转步 3.

步 3 $d^k := -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$.

步 4 进行一维精确搜索，即求步长 α_k，满足

$$f(x^k + \alpha_k d^k) = \min\{f(x^k + \alpha d^k) \mid \alpha \geqslant 0\},$$

$$x^{k+1} = x^k + \alpha_k d^k, k := k + 1. \text{转步 } 2.$$

算法分析 只要$\nabla^2 f(x^k)$正定，这种算法一般具有全局收敛性和超线性的收敛速度. 但该算法也有明显的缺点，即$\nabla^2 f(x^k)$常常是不正定的，即使正定，在每一步迭代过程中，$(\nabla^2 f(x^k))^{-1}$的计算量也太大. 为避免这些缺点，人们提出了拟牛顿算法. 下面介绍 DFP 算法(拟牛顿算法的一种).

1.4 DFP 算法

算法 4(**DFP 算法**)

步1 取 $x^1 \in \mathbf{R}^n$，允许误差 $\varepsilon > 0$.

步2 取矩阵 $H_1 := I_n, k := 1$.

步3 $d^k := -H_k \nabla f(x^k)$，进行一维精确搜索，即求步长 α_k 满足

$$f(x^k + \alpha_k d^k) = \min\{f(x^k + \alpha d^k) \mid \alpha \geqslant 0\}.$$

步4 $x^{k+1} := x^k + \alpha_k d^k$.

步5 若 $\| \nabla f(x^{k+1}) \| \leqslant \varepsilon$，算法终止，$x^{k+1}$ 作为近似最优解；否则转步6.

步6 若 $k = n$，则令 $x^1 = x^{k+1}$，返回步2；否则转步7.

步7 令 $p_k := x^{k+1} - x^k, q_k = \nabla f(x^{k+1}) - \nabla f(x^k)$,

$$H_{k+1} := H_k + \frac{p_k p_k^{\mathrm{T}}}{p_k^{\mathrm{T}} q_k} - \frac{H_k q_k q_k^{\mathrm{T}} H_k}{q_k^{\mathrm{T}} H_k q_k},$$

$k := k + 1$，返回步3.

例3 用 DFP 算法求函数

$$f(x) = 2x_1^2 + x_2^2 - 4x_1 + 2$$

的极小点.

解 取

$$x^1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad H_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

通过计算得

$$\nabla f(x^1) = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \quad d^1 = \begin{bmatrix} -4 \\ -2 \end{bmatrix}, \quad \alpha_1 = 5/18, x^2 = \begin{bmatrix} -4/9 \\ 8/9 \end{bmatrix}.$$

再进行第二次迭代，由计算得

$$p_1 = \alpha_1 d^1 = \begin{bmatrix} -10/9 \\ -5/9 \end{bmatrix}, \quad q_1 = \begin{bmatrix} -40/9 \\ -10/9 \end{bmatrix},$$

$$H_2 = 1/306 \begin{bmatrix} 86 & -38 \\ -38 & 305 \end{bmatrix},$$

$$d^2 := -H_2 \nabla f(x^2) = 12/51 \begin{bmatrix} 1 \\ -4 \end{bmatrix}.$$

从 x^2 出发沿 d^2 方向进行搜索，得 $\alpha_2 = 17/36, x^3 = x^2 + \alpha_2 d^2 = (1,0)^{\mathrm{T}}$. 这时 $\nabla f(x^3) = (0,0)^{\mathrm{T}}$，从而 $x^3 = (1,0)^{\mathrm{T}}$ 是最优解.

算法分析 拟牛顿算法因其修正矩阵的不同而形成多种算法. DFP 算法是其中主要的一种，它具有很多重要性质. 例如，对于二次函数，该算法经有限步迭代后即可求出极小点；对于凸函数，该算法具有全局收敛性；对于一般目标函数，该算法具有超线性的收敛速度.

1.5 共轭梯度算法

1.5.1 算法简介

共轭梯度算法的一般迭代程序是

$$x^{k+1} := x^k + \alpha_k d^k,$$

其中

$$d^k = \begin{cases} -\nabla f(x^k), & k = 1, \\ -\nabla f(x^k) + \beta_k d^{k-1}, & k \geq 2. \end{cases}$$

当目标函数是二次凸函数时,即 $f(x) = \frac{1}{2}x^{\mathrm{T}}Ax + b^{\mathrm{T}}x + c$,(其中 A 是对称正定矩阵),参数 β_k 按下面的公式计算:

$$\beta_k = \frac{\nabla f(x^k)^{\mathrm{T}} A\, d^{k-1}}{(d^{k-1})^{\mathrm{T}} A\, d^{k-1}}.$$

在步长 α_k 按一维精确搜索计算的情况下,由此得到的 $d^1, d^2, \cdots, d^k$ 关于 A 共轭,即 $(d^i)^{\mathrm{T}} A d^j = 0 (i \neq j)$. 可以证明至多迭代 n 次即可求出极小点.

将共轭梯度法推广到非二次函数时,参数 β_k 有多种计算公式. 因计算公式不同,构成了不同的共轭梯度算法,下面介绍其中主要的两种.

1.5.2　FR 方法和 PRP 方法

算法 5(FR 方法)

步 1　取初始点 x^1, $k = 1$, 允许误差 $\varepsilon > 0$.

步 2　计算 $\nabla f(x^k)$.

步 3　若 $\|\nabla f(x^k)\| \leq \varepsilon$, 则算法终止, x^k 为近似极小点; 否则, 令 $d^k = -\nabla f(x^k) + \beta_{k-1} d^{k-1}$, 其中

$$\beta_{k-1} = \begin{cases} 0, & \text{当 } k = 1 \text{ 时}; \\ \dfrac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^{k-1})\|^2}, & \text{当 } k > 1 \text{ 时}. \end{cases}$$

步 4　进行一维搜索,即求步长 α_k 满足

$$f(x^k + \alpha_k d^k) = \min\{f(x^k + \alpha d^k) \mid \alpha \geq 0\}.$$

步 5　$x^{k+1} = x^k + \alpha_k d^k$, $k = k + 1$, 返回步 2.

在上述迭代算法中,将 β_{k-1} 的计算公式替换成

$$\beta_{k-1} = \begin{cases} 0, & \text{当 } k = 1 \text{ 时}; \\ \dfrac{\nabla f(x^k)^{\mathrm{T}}(\nabla f(x^k) - \nabla f(x^{k-1}))}{\|\nabla f(x^{k-1})\|^2}, & \text{当 } k > 1 \text{ 时}, \end{cases}$$

就得到 **PRP 方法**.

例 4　用 FR 方法求

$$f(x_1, x_2) = (x_1 - 2)^2 + (x_1 - 2x_2)^2$$

的极小点.

解　取初始点 $x^1 = (0,3)^{\mathrm{T}}$. $d^1 = -\nabla f(x^1) = (44, -24)^{\mathrm{T}}$.

对 $f(x^1 + \alpha d^1)$ 关于 $\alpha \geq 0$ 进行一维搜索得 $\alpha_1 = 0.062$, 从而

$$x^2 = x^1 + \alpha_1 d^1 = (2.70, 1.51)^{\mathrm{T}}.$$

转入下一迭代过程.

对 $x^2 = (2.70,1.51)^T$ 进行迭代. 经计算得

$$\nabla f(x^2) = (0.73,1.28)^T,$$
$$\beta_1 = \| \nabla f(x^2) \|^2 / \| \nabla f(x^1) \|^2 = 0.00086,$$
$$d^2 = -\nabla f(x^2) + \beta_1 d^1 = (-0.09, -1.30)^T.$$

对 $f(x^2 + \alpha d^2)$ 关于 $\alpha \geqslant 0$ 进行一维搜索得 $\alpha_2 = 0.23$, 故

$$x^3 = x^2 + \alpha_2 d^2 = (2.54,1.21)^T.$$

对 x^3 重复上述迭代过程. 当迭代到一定程度时, $\| \nabla f(x^k) \|$ 小于允许误差, x^k 即逼近极小点 $(2,1)^T$. 在此不再赘述.

算法分析 可以证明 FR 方法在一般情况下具有全局收敛性, 而 PRP 方法却不能保证具有此性质. 鲍威尔(Powell) 曾举出一个有趣的例子, 在精确搜索下, PRP 方法产生的点列 $\{x^k\}$ 在6个点中间循环, 这6个点中任何一个均不是目标函数的稳定点. 但 PRP 方法的数值计算效果要比 FR 方法好得多. 因此, 在实际计算中, 人们常常将这两种方法结合使用.

2 约束最优化问题的最优性条件

2.1 约束最优化问题

2.1.1 基本概念

约束优化问题可以表示为

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & c_i(x) = 0, i \in E = \{1,2,\cdots,P\}; \\ & c_i(x) \geqslant 0, i \in I = \{p+1,p+2,\cdots,m\}. \end{aligned} \tag{2-1}$$

其中 $f(x)$ 与 $c_i(x)(i \in E \cup I)$ 是开集 $D \subset \mathbf{R}^n$ 上的连续可微函数. 如果 $I = \varnothing$, 则称问题(2-1) 式为等式约束优化问题. 若 $c_i(x)(i \in E \cup I)$ 为线性函数, 则称问题(2-1) 式为**线性约束优化问题**. 特别地, 如果目标函数 $f(x)$ 是二次函数, 则称其为**二次规划问题**.

1. 可行域与可行点

集合

$$\Omega = \{x : c_i(x) = 0, i \in E, c_i(x) \geqslant 0, i \in I\}$$

称为(2-1) 式的可行域. 如果 $x \in \Omega$, 则称 x 是(2-1) 式的**可行点**.

2. 全局极小点与全局严格极小点

设 $x^* \in \Omega$, 如果 $\forall x \in \Omega$, 均有

$$f(x) \geqslant f(x^*)$$

成立, 则称 x^* 为问题(2-1) 的全局极小点. 如果 $\forall x \in \Omega \setminus \{x^*\}$, 均有

$$f(x) > f(x^*)$$

成立,则称 x^* 为问题(2-1) 的全局严格极小点.

3. 局部极小点

设 $x^* \in \Omega$,如果对某一 $\delta > 0$,使 $\forall x \in \Omega \cap B(x^*,\delta)$,均有

$$f(x) \geqslant f(x^*)$$

成立,则称 x^* 为问题(2-1) 的局部极小点.其中,

$$B(x^*,\delta) = \{x : \| x - x^* \| \leqslant \delta\}.$$

4. 可行方向

设 $x^* \in \Omega, d \in \mathbf{R}^n, d \neq 0$.如果存在 $\delta > 0$,使 $\forall t \in (0,\delta]$,均有 $x^* + td \in \Omega$,则称 d 为 Ω 在 x^* 处的**可行方向**.

5. 有效约束与有效集

设 $\bar{x} \in \mathbf{R}^n$,称下标属于集合 $\mathscr{A}(\bar{x}) = \{i \mid i \in E \cup I, c_i(\bar{x}) = 0\}$ 的约束为 $\bar{x}$ 处的**有效约束**. $\mathscr{A}$ 称为 $(\bar{x})$ 处的有效集.

记

$$I(\bar{x}) = \mathscr{A} \cap I = \{i \mid i \in I, c_i(\bar{x}) = 0\},$$

称 $I(\bar{x})$ 为 $\bar{x}$ 处的不等式约束有效集.

2.1.2　约束规范

1. M-F 约束规范

设 $x^* \in \Omega$,如果梯度 $\nabla c_i(x^*)(i \in E)$ 线性无关,并且存在向量 s,使

$$s^{\mathrm{T}} \nabla c_i(x^*) = 0, \quad i \in E;$$

$$s^{\mathrm{T}} \nabla c_i(x^*) > 0, \quad i \in I(x^*)$$

成立,则称问题(2-1) 的约束函数在 x^* 处满足 **M-F 约束规范**.称 s 为 x^* 处的**强容许方向**.

2. 凸规划问题

所谓凸规划问题是指这样一类特殊的约束优化问题,即

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & c_i(x) = 0, i \in E = \{1,2,\cdots,P\}; \\ & c_i(x) \geqslant 0, i \in I = \{p+1,p+2,\cdots,m\}. \end{aligned} \tag{2-2}$$

这里等式约束函数 $c_i(x)(i \in E)$ 为 x 的线性函数,$f(x)$ 和不等式约束函数 $c_i(x)(i \in I)$ 则分别是开凸集 $D \subset \mathbf{R}^n$ 上的凸函数和凹函数.

3. 斯莱特(Slater) 约束规范

对于凸规划问题(2-2),如果有可行点严格满足不等式约束,即

$$S^0 = \{x \mid c_i(x) = 0, (i \in E), c_i(x) > 0 (i \in I)\} \neq \varnothing,$$

则称约束函数满足**斯莱特约束规范**.

容易证明,对于凸规划问题(2-2),若斯莱特约束规范满足,则 M-F 约束规范也满足.

2.2　最优性一阶条件

2.2.1　最优性一阶条件与 K-T 点

定理 1　**(库恩 - 塔克(Kuhn-Tucker)最优性一阶必要条件)**　设 $f(\boldsymbol{x}), c_i(\boldsymbol{x})(i \in E \cup I)$ 在开集 $D \supset \Omega$ 上连续可微.又设 $\boldsymbol{x}^*$ 为问题(2-1)的一个局部极小点,并设在 $\boldsymbol{x}^*$ 处,约束函数满足 M-F 约束规范,则必存在 $\lambda_i \in \mathbf{R}(i \in E \cup I)$,使得

$$\nabla f(\boldsymbol{x}^*) = \sum_{i \in E \cup I} \lambda_i \nabla c_i(\boldsymbol{x}^*),$$
$$\lambda_i \geqslant 0, \quad i \in I;$$
$$\lambda_i c_i(\boldsymbol{x}^*) = 0, \quad i \in I.$$

设 $\boldsymbol{x}^* \in \Omega$. 如果存在 $\lambda_i \in \mathbf{R}(i \in E \cup I)$,使得

$$\nabla f(\boldsymbol{x}^*) = \sum_{i \in E \cup I} \lambda_i \nabla c_i(\boldsymbol{x}^*),$$
$$\lambda_i \geqslant 0, i \in I;$$
$$\lambda_i c_i(\boldsymbol{x}^*) = 0, i \in I$$

成立,则称 $\boldsymbol{x}^*$ 为问题(2-1)的库恩 - 塔克点,简称 **K-T 点**. $\lambda_i \in \mathbf{R}(i \in E \cup I)$ 称为 $\boldsymbol{x}^*$ 处的拉格朗日(Lagrange)乘子. $\boldsymbol{x}^*$ 与 λ 称为问题(2-1)的一个 K-T 对,其中 λ 中的分量由 $\lambda_i(i \in E \cup I)$ 构成.

2.2.2　凸规划问题的最优性一阶条件

定理 2　对于凸规划问题(2-2),如果其约束函数满足斯莱特约束规范,则凸规划问题(2-2)的极小点 $\boldsymbol{x}^*$ 必是问题(2-2)的 K-T 点.

定理 3　对于凸规划问题(2-2),其 K-T 点为凸规划问题(2-2)的全局极小点.

推论　对于凸规划问题

$$\begin{aligned} \min \quad & f(\boldsymbol{x}), \\ \text{s.t.} \quad & c_i(\boldsymbol{x}) = 0, i \in E = \{1,2,\cdots,P\}, \end{aligned} \tag{2-3}$$

可行域 $\Omega = \{\boldsymbol{x} \mid c_i(\boldsymbol{x}) = 0, i \in E\} \neq \varnothing$,设 $\boldsymbol{x}^* \in \Omega$,则 $\boldsymbol{x}^*$ 为(2-3)式的 K-T 点,当且仅当 $\boldsymbol{x}^*$ 为(2-3)式的全局极小点.

2.3　最优性二阶条件

2.3.1　最优性二阶必要条件

下面讨论问题(2-1)的 K-T 点与其局部极小点之间的关系.有例子说明,K-T 点并非该问题的局部极小点.

例 1　$\min \quad x_1^2 + x_2^2,$

$\text{s.t.} \quad x_1^2 + x_2^2 - 1 \geqslant 0.$

易于验证，$x^* = (0,1)^T$ 为上述最优化问题的一个 K-T 点，并且 $\nabla^2 f(x^*)$ 正定，但 x^* 并非该问题的局部极小点.

定理 4(最优性二阶必要条件)　设 $f(x)$ 和 $c_i(x)(i \in E \cup I)$ 在开集 $D \supset \Omega$ 上二阶连续可微，并且约束函数全为线性函数或者 $\nabla c_i(x^*)(i \in E \cup I(x^*))$ 线性无关($x^* \in \Omega$). 又设 x^* 与 λ^* 是问题(2-1)的一个 K-T 对. 若 x^* 为问题(2-1)的局部极小点，则

$$\forall s \in S = \{s \mid s^T \nabla c_i(x^*) = 0 \quad (i \in E \cup I(x^*))\}$$

均有 $s^T \nabla_x^2 L(x^*, \lambda^*) s \geq 0$. 这里

$$L(x,\lambda) = f(x) - \sum_{i \in E \cup I(x^*)} \lambda_i c_i(x).$$

2.3.2　最优性二阶充分条件

定理 5(最优性二阶充分条件)　设 $f(x)$ 和 $c_i(x)(i \in E \cup I)$ 在开集 $D \supset \Omega$ 上二阶连续可微，并设 x^*, λ^* 是问题(2-1)的一个 K-T 对. 如果对满足

$$s^T \nabla c_i(x^*) = 0, \qquad i \in E;$$
$$s^T \nabla c_i(x^*) = 0, \qquad i \in I(x^*) \text{ 且 } \lambda_i^* > 0;$$
$$s^T \nabla c_i(x^*) \geq 0, \qquad i \in I(x^*) \text{ 且 } \lambda_i^* = 0$$

的任一非零向量 s，均有 $s^T \nabla_x^2 L(x^*, \lambda^*) s > 0$，则 x^* 为问题(2-1)的一个严格局部极小点.

雅可比唯一性条件　设 $f(x)$ 和 $c_i(x)(i \in E \cup I)$ 在开集 $D \supset \Omega$ 上二阶连续可微，x^* 和 λ^* 为(2-1)式的一个 K-T 对. 如果在 x^* 处满足

(1) $\nabla c_i(x^*)(i \in E \cup I(x^*))$ 线性无关；

(2) (2-1)式在 x^* 处具有严格互补性，即 $\forall i \in I(x^*)$，有 $\lambda_i^* > 0$；

(3) 对任意

$$s \in \{s \mid s^T \nabla c_i(x^*) = 0 (i \in E \cup I(x^*)\}, s \neq 0,$$

有

$$s^T \nabla_x^2 L(x^*, \lambda^*) s > 0,$$

则称(2-1)式在 x^* 处满足雅可比唯一性条件.

定理 6　设 $f(x)$ 和 $c_i(x)(i \in E \cup I)$ 在开集 $D \supset \Omega$ 上二阶连续可微，x^* 与 λ^* 为(2-1)式上的一个 K-T 对，并且(2-1)式在 x^* 处满足雅可比唯一性条件，则 K-T 关系式

$$\begin{cases} \nabla_x L(x,\lambda) = 0, \\ c_i(x) = 0, i \in E, \\ \lambda_i c_i(x) = 0, i \in I, \end{cases}$$

在 (x^*, λ^*) 处的雅可比矩阵 $J(x^*, \lambda^*)$ 为非奇异，而且 x^* 为(2-1)式的一个严格局部极小点.

3 罚函数法与有效集方法

3.1 罚函数法

罚函数法是通过求解一个或多个罚函数的极小点来求解约束优化问题的方法.它的基本思想是将约束问题无约束化.下面主要介绍罚函数法中的外点法和内点法.

3.1.1 外点法

罚函数　对于约束优化问题(2-1)式,称函数 $F(\boldsymbol{x},\sigma)=f(\boldsymbol{x})+\sigma P(\boldsymbol{x})$ 为罚函数,其中

$$P(\boldsymbol{x})=\sum_{i\in E}[c_i(\boldsymbol{x})]^2+\sum_{i\in I}[c_i(\boldsymbol{x})_-]^2,$$
$$c_i(\boldsymbol{x})_-=\min\{0,c_i(\boldsymbol{x})\}.$$

算法 1(外点法)

步 1　给定初始点 $\boldsymbol{x}^0$,初始惩罚因子 $\sigma_1>0$,放大系数 $c>1$,允许误差 $\varepsilon>0$, $k=1$.

步 2　以 $\boldsymbol{x}^{k-1}$ 为初始点,求解无约束问题.

$$\min\quad f(\boldsymbol{x})+\sigma_k P(\boldsymbol{x}),$$

设极小点为 $\boldsymbol{x}^k$.

步 3　若 $\sigma_k P(\boldsymbol{x}^k)<\varepsilon$,则算法终止,$\boldsymbol{x}^k$ 作为近似极小点;否则,令 $\sigma_{k+1}=c\,\sigma_k$, $k:=k+1$,转步 2.

例 1　求下述优化问题的极小点:

$$\begin{aligned}&\min\quad (x_1-1)^2+x_2^2,\\&\text{s.t.}\quad x_2-1\geqslant 0.\end{aligned}$$

解　定义罚函数

$$\begin{aligned}F(\boldsymbol{x},\sigma)&=(x_1-1)^2+x_2^2+\sigma[\min(0,x_2-1)]^2\\&=\begin{cases}(x_1-1)^2+x_2^2, & \text{当 } x_2\geqslant 1\text{ 时};\\(x_1-1)^2+x_2^2+\sigma(x_2-1)^2, & \text{当 } x_2<1\text{ 时}.\end{cases}\end{aligned}$$

于是

$$\frac{\partial F(\boldsymbol{x},\sigma)}{\partial x_1}=2(x_1-1),$$

$$\frac{\partial F(\boldsymbol{x},\sigma)}{\partial x_2}=\begin{cases}2x_2, & \text{当 } x_2<1\text{ 时};\\2x_2+2\sigma(x_2-1), & \text{当 } x_2\geqslant 1\text{ 时}.\end{cases}$$

令

$$\frac{\partial F}{\partial x_1}=\frac{\partial F}{\partial x_2}=0,$$

得
$$\bar{x}_\sigma = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \dfrac{\sigma}{1+\sigma} \end{bmatrix}.$$

令 $\sigma \to +\infty$ 得　$\bar{x}_\sigma \xrightarrow[\sigma \to +\infty]{} \bar{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. $\bar{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ 即为优化问题的极小点.

3.1.2　内点法

罚函数的内点法适用于只有不等式约束的优化问题

$$\begin{aligned} \min \quad & f(x), \\ \text{s.t.} \quad & c_i(x) \geqslant 0, i \in I. \end{aligned} \tag{3-1}$$

其中 $f(x), c_i(x)(i \in I)$ 是开集 $D \subset \mathbf{R}^n$ 上的连续可微函数.

(1) 障碍函数　称函数 $G(x,\sigma) = f(x) + \frac{1}{\sigma}B(x)$ 为**障碍函数**.其中 $B(x)$ 是开集 $D \subset \mathbf{R}^n$ 上的连续可微函数,且当 x 趋于可行域 Ω 的边界时,$B(x) \to +\infty$.

(2) 倒数罚函数　函数 $G(x,\sigma) = f(x) + \frac{1}{\sigma}\sum_{i \in I} 1/c_i(x)$ 称为**倒数罚函数**.

(3) 对数罚函数　函数 $G(x,\sigma) = f(x) - \frac{1}{\sigma}\sum_{i \in I} \log c_i(x)$ 称为**对数罚函数**.

显然,倒数罚函数和对数罚函数均是障碍函数.其中 $B(x) = \sum_{i \in I} 1/c_i(x)$ 和 $B(x) = -\sum_{i \in I} \log c_i(x)$.

算法2(内点法)

步1　给定初始点 $x^0 \in \text{int}\Omega$,允许误差 $\varepsilon > 0, \sigma_1 > 0, c > 1, k = 1$.

步2　利用初始值 x^{k-1},求解优化问题

$$\min \quad G(x,\sigma_k),$$

求得极小点 x^k 这里 $G(x,\sigma_k)$ 为倒数罚函数或对数罚函数.

步3　若 $1/\sigma_k B_k < \varepsilon$,算法终止,$x^k$ 作为近似极小点;否则,$\sigma_{k+1} = c\sigma_k, k = k + 1$,返回步2.

例2　利用罚函数的内点法求解下述优化问题:

$$\begin{aligned} \min \quad & \frac{1}{12}(x_1+3)^3 + x_2; \\ \text{s.t.} \quad & x_1 - 1 \geqslant 0, \\ & x_2 \geqslant 0. \end{aligned}$$

解　取罚函数

$G(x,\sigma_k) = \frac{1}{12}(x_1+3)^3 + x_2 + \frac{1}{\sigma_k}\left(\frac{1}{x_1-1} + 1/x_2\right)$,用解析法求优化问题

$$\min \quad G(x,\sigma_k)$$

的极小点.令

$$\frac{\partial G(x,\sigma_k)}{\partial x_1} = \frac{\partial G(x,\sigma_k)}{\partial x_2} = 0,$$

得

$$\bar{x}(\sigma_k) = \begin{bmatrix} \sqrt{1 + \dfrac{2}{\sqrt{\sigma_k}}} \\ \dfrac{1}{\sqrt{\sigma_k}} \end{bmatrix}.$$

令 $\sigma_k \to +\infty$，$\bar{x}(\sigma_k) \to \bar{x} = (1,0)^T$，$\bar{x} = (1,0)^T$ 即为原优化问题的极小点.

算法分析 罚函数的外点法和内点法均采用无约束极小化技巧.其方法简单,实用方便,并能用来求解导数不存在的情况,而且具有很好的收敛性.但这种罚函数法的缺点是,当乘子 σ_k 无限增大时,求罚函数极小点会越来越困难.

3.2 二次规划问题的有效集方法

3.2.1 二次规划问题

1. 二次规划问题

二次规划问题是最简单的非线性规划问题,它是优化问题(2-1) 在 $f(x)$ 为二次函数,且 $c_i(x)(i \in E \cup I)$ 都是线性函数时的特殊情形,即可以写成

$$\begin{aligned} \min \quad & Q(x) = \frac{1}{2}x^T H x + g^T x; \qquad (3\text{-}2) \\ \text{s.t.} \quad & a_i^T x = b_i, \quad i \in E = \{1,2,\cdots,P\}, \\ & a_i^T x \geqslant b_i, \quad i \in I = \{p+1, p+2, \cdots, m\}. \end{aligned}$$

2. 等式约束二次规划问题的求解方法

在二次规划问题(3-2) 式中,$I = \varnothing$ 时,则称其为等式约束二次规划问题.

由于等式约束二次规划问题可以写成

$$\begin{aligned} \min \quad & Q(x) = g^T x + \frac{1}{2}x^T H x, \qquad (3\text{-}3) \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

其中 $g \in \mathbf{R}^n$,$b \in \mathbf{R}^m$,$A \in \mathbf{R}^{m\times n}$,$H \in \mathbf{R}^{n\times n}$,且 H 是对称的,不失一般性,假定 $\operatorname{rank}(A) = m$.

由于 A 行满秩,一定可以找到变量 x 的一个分解 $x = (x_B, x_N)^T$.其中 $x_B \in \mathbf{R}^m$,$x_N \in \mathbf{R}^{n-m}$,且 A 有相应的分解 $A = (A_B, A_N)$ 使得 A_B 可逆,这样优化问题(3-3) 式的约束条件可以写成

$$A_B x_B + A_N x_N = b,$$

即

$$x_B = A_B^{-1}b - A_B^{-1}A_N x_N.$$

将其代入原优化问题(3-3) 式,得到(3-3) 式的一个等价形式

$$\min_{x_N \in \mathbf{R}^{n-m}} \hat{g}_N^T x_N + \frac{1}{2}x_N^T \hat{H}_N x_N. \qquad (3\text{-}4)$$

其中

$$\hat{g}_N = g_N - A_N^T (A_B^{-1})^T g_B + [H_{NB} - A_N^T (A_B^{-1})^T H_{BB}] A_B^{-1} b;$$

$$\widehat{H}_N = H_{NN} - H_{NB}A_B^{-1}A_N - A_N^T(A_B^{-1})^T H_{BN} + A_N^T(A_B^{-1})^T H_{BB}A_B^{-1}A_N;$$

$$g = (g_B, g_N)^T,$$

和

$$H = \begin{bmatrix} H_{BB} & H_{BN} \\ H_{NB} & H_{NN} \end{bmatrix}$$

是与 $x = (x_B, x_N)^T$ 相应的分解.

这样,利用无约束优化方法求解上述优化问题(3-4)式,即可得到优化问题(3-3)式的极小点.

3.2.2 有效集方法

有效集方法是通过求解有限个等式约束二次规划问题来解决一般约束下的二次规划问题(3-2).直观上,非有效约束在解的附近不起作用,可以去掉不考虑,而不等式有效约束由于它在解处等式成立,所以可以用等式约束来替换不等式约束.

算法 3(有效集方法)

步 1 给出初始可行点 x^1, $S_1 = E \cup I(x^1)$, $k = 1$.

步 2 求解二次规划问题

$$\begin{aligned} \min \quad & g^T(x^k + d) + \frac{1}{2}(x^k + d)^T H(x^k + d), \\ \text{s.t.} \quad & a_i^T d = 0, \qquad i \in S_k, \end{aligned}$$

得 d^k 和相应的拉格朗日乘子 $\lambda_i^k (i \in S_k)$. 如果 $d^k \neq 0$,则转步 3. 如果 $\lambda_i^k \geq 0 (i \in S_k \cap I)$,则算法终止,$x^k$ 作为原规划问题的极小点;否则,取 $\lambda_{i_k}^k = \min\{\lambda_i^k : \lambda_i^k < 0, i \in S_k \cap I\}$,得 i_k,令 $S_k = S_k \setminus \{i_k\}$, $x^{k+1} = x^k$,转步 4.

步 3 求 α_k 满足

$$\alpha_k = \min\{1, \min[(b_i - a_i^T x)/(a_i^T d^k) : i \overline{\in} S_k, a_i^T d^k < 0]\}$$

$$x^{k+1} = x^k + \alpha_k d^k.$$

若 $\alpha_k = 1$,则转步 4;否则取 $j \overline{\in} S_k$,使

$$a_j^T(x^k + \alpha_k d^k) = b_j, \text{令 } S_k = S_k \cup \{j\}.$$

步 4 $S_{k+1} = S_k$, $k = k + 1$,转步 2.

在上述算法中.若 n 阶对称矩阵 H 可逆,则步 2 中的拉格朗日乘子 λ_i^k 可以用下述公式得到.由于 d^k 和 λ^k 是优化问题

$$\begin{aligned} \min \quad & g^T(x^k + d) + \frac{1}{2}(x^k + d)^T H(x^k + d), \\ \text{s.t.} \quad & a_i^T(x^k + d) = b_i, i \in S_k \end{aligned}$$

的一个 K-T 对,设上述问题的约束 $a_i^T(x^k + d) = b_i (i \in S_k)$ 的系数矩阵为 A_k,常数列向量为 b^k,则 $\lambda^k = (A_k H^{-1} A_k^T)^{-1}(A_k H^{-1} g + b^k)$.

例 3 用有效集方法求二次规划问题的极小点.

$$\begin{aligned} \min \quad & f(x) = x_1^2 - x_1 x_2 + 2x_2^2 - x_1 - 10x_2, \\ \text{s.t.} \quad & -3x_1 - 2x_2 \geq -6, x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

解　目标函数 $f(\boldsymbol{x})$ 可以写成

$$f(\boldsymbol{x}) = \frac{1}{2}(x_1, x_2)\begin{bmatrix} 2 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + (-1, -10)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

取初始可行点 $\boldsymbol{x}^1 = (0,0)^{\mathrm{T}}$，则 $S_1 = \{2,3\}$. 求解二次规划问题

$$\begin{aligned} \min \quad & d_1^2 - d_1 d_2 + 2d_2^2 - d_1 - 10d_2, \\ \text{s.t.} \quad & d_1 = 0, \qquad d_2 = 0, \end{aligned}$$

得 $\boldsymbol{d}^1 = (0,0)^{\mathrm{T}}$ 和相应的拉格朗日乘子 $\boldsymbol{\lambda}^1 = (-1, -10)^{\mathrm{T}}$. 令

$$S_2 = S_1 \setminus \{3\} = \{2\}, \boldsymbol{x}^2 = \boldsymbol{x}^1 = (0,0)^{\mathrm{T}}.$$

求解二次规划问题

$$\begin{aligned} \min \quad & d_1^2 - d_1 d_2 + 2d_2^2 - d_1 - 10d_2, \\ \text{s.t.} \quad & d_1 = 0, \end{aligned}$$

得 $\boldsymbol{d}^2 = (0,5/2)^{\mathrm{T}}$. 令

$$\alpha_2 = \min\{1,6/5\} = 1, \boldsymbol{x}^3 = \boldsymbol{x}^2 + \alpha_2 \boldsymbol{d}^2 = (0,5/2)^{\mathrm{T}}, S_3 = S_2 = \{2\},$$

再求解二次规划问题

$$\begin{aligned} \min \quad & d_1^2 - 7/2d_1 - d_1 d_2 + 2d_2^2 - 25/2, \\ \text{s.t.} \quad & d_1 = 0, \end{aligned}$$

得 $\boldsymbol{d}^3 = (0,0)^{\mathrm{T}}, \lambda^3 = -7/2$. 令

$$S_4 = S_3 \setminus \{2\} = \varnothing, \boldsymbol{x}^4 = \boldsymbol{x}^3 = (0,5/2)^{\mathrm{T}},$$

求解二次规划问题

$$\min \quad d_1^2 - 7/2d_1 - d_1 d_2 + 2d_2^2 - 25/2,$$

得 $\boldsymbol{d}^4 = (2,1/2)^{\mathrm{T}}$. 令

$$\alpha_4 = \min\{1,1/7\} = 1/7, \boldsymbol{x}^5 = \boldsymbol{x}^4 + \alpha_4 \boldsymbol{d}^4 = (2/7,18/7)^{\mathrm{T}},$$

在 $\boldsymbol{x}^5$ 点，第一个约束为积极约束，取 $S_5 = S_4 \cup \{1\} = \{1\}$，求解二次规划问题

$$\begin{aligned} \min \quad & d_1^2 - d_1 d_2 + 2d_2^2 - 3d_1 - 6/7, \\ \text{s.t.} \quad & -3d_1 - 2d_2 = 0, \end{aligned}$$

得 $\boldsymbol{d}^5 = (3/14, -9/28)^{\mathrm{T}}$. 令

$$\alpha_5 = \min\{1,8\} = 1, \boldsymbol{x}^6 = \boldsymbol{x}^5 + \alpha_5 \boldsymbol{d}^5 = (1/2,9/4)^{\mathrm{T}}, S_6 = S_5 = \{1\},$$

求解二次规划问题

$$\begin{aligned} \min \quad & d_1^2 - 9/4d_1 - 3/2d_2 - d_1 d_2 + 2d_2^2 - 55/4, \\ \text{s.t.} \quad & -3d_1 - 2d_2 = 0, \end{aligned}$$

得 $\boldsymbol{d}^6 = (0,0)^{\mathrm{T}}$ 和 $\lambda^6 = 3/4$. 故 $\boldsymbol{x}^6 = (1/2,9/4)^{\mathrm{T}}$ 是原二次规划问题的极小点.

算法分析　设点列 $\{\boldsymbol{x}_k\}$ 由有效集方法产生，如果对于任何 k 均有 $\boldsymbol{a}_i(i \in E \cup I(\boldsymbol{x}_k))$ 线性无关，则该算法必有限终止于规划问题(3-2) 的 K-T 点或者原问题(3-2) 目标函数无下界，而且若 $\boldsymbol{H}$ 为正定矩阵，该方法有限步终止时，得到的点为二次规划问题(3-2) 式的极小点. 若不能保证 $\boldsymbol{a}_i(i \in E \cup I(\boldsymbol{x}_k))$ 线性无关，则算法 3 有可能出现与线性规划问题类似的循环现象. 如果 $\boldsymbol{H}$ 不正定，则可能出现二次规划问题(3-2) 目标函数无下界.

4 可行方向法

4.1 邹迪耶克可行方向法

4.1.1 线性约束规划问题的邹迪耶克可行方向法

对于线性约束非线性规划问题

$$\begin{aligned} \min \quad & f(\boldsymbol{x}), \\ \text{s.t.} \quad & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} = b_i, i \in E = \{1,2,\cdots,p\}, \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} \geqslant b_i, i \in I = \{p+1,\cdots,m\}, \end{aligned} \tag{4-1}$$

其中 $f(\boldsymbol{x})$ 是开集 $D \subset \mathbf{R}^n$ 上的连续可微函数,有如下邹迪耶克(G.Zoutendijk)可行方向法:

算法 1(邹迪耶克可行方向法)

步 1 取初始可行点 $\boldsymbol{x}^1, k = 1$.

步 2 求出 $I(\boldsymbol{x}^k)$.

步 3 求解线性规划问题

$$\begin{aligned} \min \quad & \nabla f(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}, \\ \text{s.t.} \quad & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{d} = 0, i \in E, \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{d} \geqslant 0, i \in I(\boldsymbol{x}^k), -1 \leqslant d_j \leqslant 1, \quad j = 1,2,\cdots,n, \end{aligned}$$

得 $\boldsymbol{d}^k$.

步 4 若$\nabla f(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}^k = 0$,则算法终止,$\boldsymbol{x}^k$ 为 K-T 点;否则转步 5.

步 5 求 $\alpha_{\max}$ 满足

$$\alpha_{\max} = \begin{cases} \min\{\widehat{b}_i/\widehat{d}_i : \widehat{d}_i < 0\}, & \text{当 } \widehat{\boldsymbol{d}}^k \ngeqslant 0 \text{ 时}; \\ \infty, & \text{当 } \widehat{\boldsymbol{d}}^k \geqslant 0 \text{ 时}. \end{cases}$$

其中

$$\widehat{b}_i = b_i - \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x}^k, \quad i \in I \setminus I(\boldsymbol{x}^k);$$

$$\widehat{d}_i = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{d}^k, \quad i \in I \setminus I(\boldsymbol{x}^k).$$

$\widehat{\boldsymbol{d}}^k$ 为由 $\widehat{d}_i(i \in I \setminus I(\boldsymbol{x}^k))$,组成的向量.

步 6 求解规划问题

$$\begin{aligned} \min \quad & f(\boldsymbol{x}^k + \alpha\boldsymbol{d}^k), \\ \text{s.t.} \quad & 0 \leqslant \alpha \leqslant \alpha_{\max}, \end{aligned}$$

得 α_k,令 $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k\boldsymbol{d}^k$.

步 7 $k = k + 1$,返回步 2.

例 1 用邹迪耶克可行方向法求下述规划问题的极小点:

$$\begin{aligned}\min\quad & x_1^2+x_2^2-2x_1-4x_2+6,\\ \text{s.t.}\quad & -2x_1+x_2\geqslant -1,\\ & -x_1-x_2\geqslant -2,\ x_1\geqslant 0,\ x_2\geqslant 0.\end{aligned}$$

解 取初始可行点 $\boldsymbol{x}^1=(0,0)^{\mathrm T}$，$I(\boldsymbol{x}^1)=\{3,4\}$，解规划问题

$$\begin{aligned}\min\quad & -2d_1-4d_2,\\ \text{s.t.}\ & d_1\geqslant 0,\ d_2\geqslant 0,\ -1\leqslant d_j\leqslant 1,\quad j=1,2,\end{aligned}$$

得 $\boldsymbol{d}^1=(1,1)^{\mathrm T}$，并求得 $\alpha_{\max}=1$.

解一个变量的规划问题

$$\begin{aligned}\min\quad & 2\alpha^2-6\alpha+6,\\ \text{s.t.}\quad & 0\leqslant\alpha\leqslant 1,\end{aligned}$$

得 $\alpha_1=1$. 令 $\boldsymbol{x}^2=\boldsymbol{x}^1+\alpha_1\boldsymbol{d}^1=(1,1)^{\mathrm T}$，$I(\boldsymbol{x}^2)=\{1,2\}$.

进行第二次迭代：解规划问题

$$\begin{aligned}\min\quad & -2d_2,\\ \text{s.t.}\quad & -2d_1+d_2\geqslant 0,\\ & -d_1-d_2\geqslant 0,\ -1\leqslant d_j\leqslant 1,\quad j=1,2,\end{aligned}$$

得 $\boldsymbol{d}^2=(-1,1)^{\mathrm T}$，$\alpha_{\max}=1$.

求解规划问题

$$\begin{aligned}\min\quad & 2\alpha^2-2\alpha+2,\\ \text{s.t.}\quad & 0\leqslant\alpha\leqslant 1,\end{aligned}$$

得 $\alpha_2=1/2$. 令 $\boldsymbol{x}^3=\boldsymbol{x}^2+\alpha_2\boldsymbol{d}^2=(1/2,3/2)^{\mathrm T}$，$I(\boldsymbol{x}^3)=\{2\}$.

进行第三次迭代：解规划问题

$$\begin{aligned}\min\quad & -d_1-d_2,\\ \text{s.t.}\quad & -d_1-d_2\geqslant 0,\\ & -1\leqslant d_j\leqslant 1,\quad j=1,2,\end{aligned}$$

得 $\boldsymbol{d}^3=(0,0)^{\mathrm T}$. 由于此时 $\nabla f(\boldsymbol{x}^3)^{\mathrm T}\boldsymbol{d}^3=0$. 这样，$\boldsymbol{x}^3$ 即为原规划问题的 K-T 点. 由于此例是凸规划问题，因此 $\boldsymbol{x}^3$ 是极小点.

4.1.2 非线性约束规划问题的邹迪耶克可行方向法

对于只含不等式约束的非线性规划问题

$$\begin{aligned}\min\quad & f(\boldsymbol{x}),\\ \text{s.t.}\quad & c_i(\boldsymbol{x})\geqslant 0,\quad i\in I=\{1,2,\cdots,m\},\end{aligned}\tag{4-2}$$

其中 $f(\boldsymbol{x})$ 和 $c_i(\boldsymbol{x})(i\in I)$ 是开集 $D\subset\mathbf{R}^n$ 上的连续可微函数，有如下邹迪耶克可行方向算法.

算法 2

步 1 取初始可行点 $\boldsymbol{x}^1$，$k=1$.

步 2 求 $I(\boldsymbol{x}^k)=\{i\mid c_i(\boldsymbol{x})=0\}$.

解下述线性规划问题：

$$\begin{aligned}&\min z,\\&\text{s.t.}\ \nabla f(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}-z\leqslant 0,\\&\qquad\nabla c_i(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}+z\geqslant 0,i\in I(\boldsymbol{x}^k),\\&\qquad -1\leqslant d_j\leqslant 1,j=1,2,\cdots,n,\end{aligned}$$

得最优解$(z^k,\boldsymbol{d}^k)$,若 $z^k=0$,算法终止,$\boldsymbol{x}^k$ 为 F-J 点;否则转步 3.

步 3　求 $\alpha_{\max}$ 满足

$$\alpha_{\max}=\sup\{\alpha\mid c_i(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k)\geqslant 0,i\in I\},$$

解规划问题

$$\begin{aligned}&\min\quad f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k),\\&\text{s.t.}\quad 0\leqslant\alpha\leqslant\alpha_{\max},\end{aligned}$$

得 α_k.

步 4　$\boldsymbol{x}^{k+1}=\boldsymbol{x}^k+\alpha_k\boldsymbol{d}^k,k=k+1$,返回步 2.

由于算法2不能保证迭代产生的点列收敛于K-T点,因此人们提出了如下修正算法:

算法 3(汤普金斯 - 维艾奥(Topkis-Veinott)修正算法)

步 1　给定初始可行点 $\boldsymbol{x}^1,k=1$.

步 2　解线性规划问题

$$\begin{aligned}&\min\quad z,\\&\text{s.t.}\quad \nabla f(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}-z\leqslant 0,\\&\qquad\nabla c_i(\boldsymbol{x}^k)^{\mathrm{T}}\boldsymbol{d}+z\geqslant -c_i(\boldsymbol{x}^k),i\in I,\\&\qquad -1\leqslant d_j\leqslant 1,j=1,2,\cdots,n,\end{aligned}$$

得最优解$(z^k,\boldsymbol{d}^k)$.

步 3　若 $z^k=0$,则算法终止;否则转步 4.

步 4　求 $\alpha_{\max}$ 满足

$$\alpha_{\max}=\sup\{\alpha\mid c_i(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k)\geqslant 0,i\in I\},$$

解规划问题

$$\begin{aligned}&\min\quad f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k),\\&\text{s.t.}\quad 0\leqslant\alpha\leqslant\alpha_{\max},\end{aligned}$$

得 α_k.

步 5　$\boldsymbol{x}^{k+1}=\boldsymbol{x}^k+\alpha_k\boldsymbol{d}^k,k=k+1$,转步 2.

例 2　用汤普金斯 - 维艾奥修正算法求下述规划问题的极小点:

$$\begin{aligned}&\min\quad 2x_1^2+2x_2^2-2x_1x_2-4x_1-6x_2,\\&\text{s.t.}\quad -x_1-5x_2\geqslant -5,\\&\qquad -2x_1+x_2\geqslant 0,\\&\qquad x_1\geqslant 0,x_2\geqslant 0.\end{aligned}$$

解　取初始点 $\boldsymbol{x}^1=(0,3/4)^{\mathrm{T}}$,则$\nabla f(\boldsymbol{x}^1)=(-5.5,-3)^{\mathrm{T}}$.解线性规划问题

$$\min\quad z,$$

$$\begin{aligned} \text{s.t.}\quad & -5.5d_1-3d_2-z\leqslant 0,\\ & -d_1-5d_2+z\geqslant -15/4,\\ & d_2+z\geqslant -3/4,\\ & d_1+z\geqslant 0,\\ & d_2+z\geqslant -3/4,\\ & -1\leqslant d_j\leqslant 1,j=1,2, \end{aligned}$$

得 $\boldsymbol{d}^1=(0.7143,-0.03571)^{\mathrm{T}},z^1=-0.7143$.

$\alpha_{max}=0.84$,解规划问题

$$\begin{aligned} \min\quad & 0.972\alpha^2-4.036\alpha-3.375,\\ \text{s.t.}\quad & 0\leqslant\alpha\leqslant 0.84, \end{aligned}$$

得 $\alpha_1=0.84$,从而 $\boldsymbol{x}^2=\boldsymbol{x}^1+\alpha_1\boldsymbol{d}^1=(0.60,0.72)^{\mathrm{T}}$.

在 $\boldsymbol{x}^2$ 点,有$\nabla f(\boldsymbol{x}^2)=(-3.04,-4.32)^{\mathrm{T}}$.解线性规划问题

$$\begin{aligned} \min\quad & z,\\ \text{s.t.}\quad & -3.04d_1-4.32d_2-z\leqslant 0,\\ & -d_1-5d_2+z\geqslant -0.8,\\ & -2.4d_1+d_2+z\geqslant 0,\\ & d_1+z\geqslant -0.6,\\ & d_2+z\geqslant -0.72,\\ & -1\leqslant d_j\leqslant 1,j=1,2, \end{aligned}$$

得最优解 $\boldsymbol{d}^2=(-0.07123,0.1167)^{\mathrm{T}}$ 和 $z^2=0.2877$. 通过线性搜索得 $\alpha_{max}=1.561676$. 解线性规划问题

$$\begin{aligned} \min\quad & 0.54\alpha^2-0.2876\alpha-5.8272,\\ \text{s.t.}\quad & 0\leqslant\alpha\leqslant 1.561676, \end{aligned}$$

得 $\alpha_2=1.561676$,从而 $\boldsymbol{x}^3=\boldsymbol{x}^2+\alpha_2\boldsymbol{d}^2=(0.4888,0.9022)^{\mathrm{T}}$.

重复上述迭代过程,直到第五次迭代结束,得到点 $\boldsymbol{x}^6=(0.6548,0.8575)^{\mathrm{T}},z^5=-0.0303$,目标函数值 $f(\boldsymbol{x}^6)=-6.5590$ 与目标函数最优值 $f(\boldsymbol{x}^*)=-6.613086$ 比较接近.

算法分析 汤普金斯-维艾奥修正算法可以保证迭代产生的点列收敛于 F-J 点.

4.2 既约梯度法

4.2.1 简易既约梯度法

1. 线性约束最优化问题

称一个最优化问题是线性约束最优化问题,是指它可以表示成如下形式:

$$\begin{aligned} \min\quad & f(\boldsymbol{x}),\\ \text{s.t.}\quad & \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b},\boldsymbol{x}\geqslant 0. \end{aligned}\tag{4-3}$$

其中系数矩阵 $A \in \mathbf{R}^{m\times n}$，$b \in \mathbf{R}^m$，$m \leqslant n$，$f(x)$ 是开集 $D \subset \mathbf{R}^n$ 上的连续可微函数.

2. 基与基变量

将线性约束最优化问题(4-3)式中的函数矩阵 A 分块为 $A = (B, N)$，其中 $B \in \mathbf{R}^{m\times m}$，则 x 有相应的分块 $x = \begin{bmatrix} x_B \\ x_N \end{bmatrix}$. x_B 为 $Ax = b$ 中 B 中列所对应的变量. 若 B 为非奇异的，则 B 称为线性约束最优化问题(4-3)式的一个基，x_B 称为对应于基 B 的**基变量**，x_N 称为非基变量. 此时有 $x_B = B^{-1}b - B^{-1}Nx_N$.

用 J 表示基变量的下标集合，$\bar{J}$ 表示非基变量的下标集合，下面给出**伍尔夫(Wolfe)简易既约梯度算法**.

算法 4(沃尔夫既约梯度法)

步 1　取初始可行解 x^0，将 x^0 分块成 $\begin{bmatrix} x_B^0 \\ x_N^0 \end{bmatrix}$，其中 $x_B^0 > 0$ 为基变量. 对应地，A 分块成 (B, N). 给定允许误差 $\varepsilon > 0$，$k := 0$.

步 2　计算：

$$r(x_N^k) = \nabla_{x_N} f(x_B(x_N^k), x_N^k) - (B^{-1}N)^{\mathrm{T}} \nabla_{x_B} f(x_B(x_N^k), x_N^k).$$

由算式

$$p_j^k = \begin{cases} 0, & \text{当 } x_j^k = 0 \text{ 且 } r_j(x_N^k) > 0,(j \in \bar{J}); \\ -r_j(x_N^k), & \text{其余情形}(j \in \bar{J}), \end{cases}$$

得 p_N^k.

$$p_B^k = -B^{-1}Np_N^k, p^k = \begin{bmatrix} p_B^k \\ p_N^k \end{bmatrix}.$$

步 3　若 $\| p^k \| \leqslant \varepsilon$，则以 x^k 作为原问题的近似最优解，算法终止；否则令

$$\alpha_{\max} = \min\{-x_j^k/p_j^k \mid p_j^k < 0\}，\text{转步 4}.$$

步 4　求 α_k 使其满足

$$f(x^k + \alpha_k p^k) = \min\{f(x^k + \alpha p^k) \mid 0 \leqslant \alpha \leqslant \alpha_{\max}\},$$

$$x^{k+1} = x^k + \alpha_k p^k.$$

步 5　若 $\| x^{k+1} - x^k \| \leqslant \varepsilon$，以 x^{k+1} 作为近似最优解，算法终止；否则转步 6.

步 6　若 $x_B^{k+1} > 0$，则基变量不变，$k = k+1$，返回步 2；若存在 $j_0 \in J$，使 $x_{j_0}^{k+1} = 0$，则通过转轴运算将 $x_{j_0}^{k+1}$ 换出基，而以 $x_j^{k+1}(j \in \bar{J})$ 中最大的分量换入基，构成新的基变量 x_B^{k+1} 与非基变量 x_N^{k+1}，$k = k+1$，返回步 2.

例 3　用简易既约梯度法解下述规划问题：

$$\min \quad (x_1 - 3)^2(4 - x_2),$$

$$\text{s.t.} \quad x_1 + x_2 \leqslant 3, 0 \leqslant x_1 \leqslant 2,$$

$$0 \leqslant x_2 \leqslant 2.$$

解　首先引入松弛变量 x_3, x_4, x_5，将原问题化成如下标准形式的线性约束最优化问题：

$$\begin{aligned}
\min \quad & (x_1-3)^2(4-x_2),\\
\text{s.t.} \quad & x_1+x_2+x_3=3,\\
& x_1 \qquad\; +x_4=2,\\
& \qquad x_2+x_5=2,\\
& x_j \geqslant 0, j=1,2,\cdots,5.
\end{aligned}$$

取初始可行解 $\boldsymbol{x}^0=(0,2,1.8,1,1.8,0.2)^{\mathrm{T}}$,取 x_1,x_2,x_3 为基变量,则

$$\boldsymbol{B}=\begin{bmatrix}1&1&1\\1&0&0\\0&1&0\end{bmatrix},\quad \boldsymbol{N}=\begin{bmatrix}0&0\\1&0\\0&1\end{bmatrix}.$$

故

$$\boldsymbol{x}_B=\boldsymbol{B}^{-1}\boldsymbol{b}-\boldsymbol{B}^{-1}\boldsymbol{N}\boldsymbol{x}_N=\begin{bmatrix}3-x_4\\2-x_5\\-1+x_4+x_5\end{bmatrix}.$$

通过计算得

$$\boldsymbol{r}(\boldsymbol{x}_N^0)=\begin{bmatrix}12.32\\7.84\end{bmatrix},$$

$$\boldsymbol{p}^0=(12.32,7.84,-20.16,-12.32,-7.84)^{\mathrm{T}},$$

$$\alpha_{\max}=\min\{1/20.16,1.8/12.32,0.2/7.84\}=0.2/7.84.$$

解优化问题

$$\begin{aligned}
\min \quad & f(\boldsymbol{x}^0+\alpha\boldsymbol{p}^0)=(0.2+12.32\alpha-3)^2(4-1.8-7.84\alpha),\\
\text{s.t.} \quad & 0\leqslant 0\leqslant 0.2/7.84,
\end{aligned}$$

得 $\alpha_0=\alpha_{\max}=0.2/7.84$,从而

$$\boldsymbol{x}^1=\boldsymbol{x}^0+\alpha_0\boldsymbol{p}^0=(0.512,2,0.488,1.488,0)^{\mathrm{T}}.$$

由于 $\boldsymbol{x}_B^1=(0,512,2,0.488)^{\mathrm{T}}>0$,故基不变,重复上述计算过程得

$$\boldsymbol{r}(\boldsymbol{x}_N^1)=\begin{bmatrix}9.95\\6.18\end{bmatrix}.$$

$$\boldsymbol{p}^1=(9.95,0,-9.95,-9.95,0)^{\mathrm{T}},$$

$$\alpha_1=\alpha_{\max}=0.488/9.95,$$

$$\boldsymbol{x}^2=\boldsymbol{x}^1+\alpha_1\boldsymbol{p}^1=(1,2,0,1,0)^{\mathrm{T}}.$$

由于 $\boldsymbol{x}_B^2\not> 0$,必须进行换基;取 x_3 出基,x_4 入基,这时

$$\boldsymbol{B}=\begin{bmatrix}1&1&0\\1&0&1\\0&1&0\end{bmatrix},\quad \boldsymbol{N}=\begin{bmatrix}1&0\\0&0\\0&1\end{bmatrix}.$$

由计算得

$$\boldsymbol{r}(\boldsymbol{x}_N^2)=\begin{bmatrix}8\\-4\end{bmatrix},\quad \boldsymbol{p}^2=(4,-4,0,-4,4)^{\mathrm{T}},$$

$$\alpha_2=\alpha_{\max}=1/4,\boldsymbol{x}^3=\boldsymbol{x}^2+\alpha_2\boldsymbol{p}^2=(2,1,0,0,1)^{\mathrm{T}}.$$

由于 $\boldsymbol{x}_B^3\not> 0$,将 x_4 换出基,而将 x_5 换入基,这时,

$$B=\begin{bmatrix}1&1&0\\1&0&0\\0&1&1\end{bmatrix},\quad N=\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}.$$

由计算得

$$r(x_N^3)=\begin{bmatrix}1\\5\end{bmatrix},p^3=(0,0,0,0,0)^{\mathrm{T}},\text{从而 } x^3 \text{ 为原规划问题的最优解}.$$

4.2.2 广义既约梯度法

阿巴迪(Abadie)和卡彭泰尔(Carpentier)将沃尔夫简易既约梯度法推广到非线性约束最优化问题而得到**广义既约梯度法(GRG 方法)**.假设非线性约束最优化问题为如下形式:

$$\begin{aligned}\min\quad & f(x),\\ \text{s.t.}\quad & h_i(x)=0,i\in E=\{1,2,\cdots,m\},\\ & \alpha\leqslant x\leqslant\beta,\end{aligned}\tag{4-4}$$

其中 $f(x)$ 和 $h_i(x)(i\in E)$ 都是开集 $D\subset\mathbf{R}^n$ 上的连续可微函数,α,β 为 n 维列向量,$m\leqslant n$.

记 $H(x)=(h_1(x),h_2(x),\cdots,h_m(x))^{\mathrm{T}}$,把 $H(x)$ 的雅可比矩阵

$$\frac{\partial H}{\partial x}=\begin{bmatrix}\dfrac{\partial h_1}{\partial x_1}&\cdots&\dfrac{\partial h_1}{\partial x_n}\\ \vdots&&\vdots\\ \dfrac{\partial h_m}{\partial x_1}&\cdots&\dfrac{\partial h_m}{\partial x_n}\end{bmatrix}$$

分块成

$$\frac{\partial H}{\partial x}=\left(\frac{\partial H}{\partial x_B},\frac{\partial H}{\partial x_N}\right),\quad \text{其中 } x=\begin{bmatrix}x_B\\x_N\end{bmatrix}.$$

假设 $\dfrac{\partial H}{\partial x_B}$ 非奇异,则 x_B 称为基变量,x_N 称为非基变量,J 和 $\bar{J}$ 分别表示基变量与非基变量的下标集.下面给出针对非线性规划问题(4-4)式的广义既约梯度算法.

算法5(GRG 方法)

步1 给定初始可行点 $x^0=\begin{bmatrix}x_B^0\\x_N^0\end{bmatrix}$,允许误差 $\varepsilon_1,\varepsilon_2>0$,正整数 $M>0,k=0$.

步2 计算

$$r(x_N^k)=\nabla_{x_N}f(x^k)-\left[\left(\frac{\partial H(x^k)}{\partial x_B}\right)^{-1}\frac{\partial H(x^k)}{\partial x_N}\right]^{\mathrm{T}}\nabla_{x_B}f(x^k).$$

由算式

$$p_j^k=\begin{cases}0, & x_j^k=\alpha_j,\text{且 } r_j(x_N^k)>0,(j\in \bar{J})\text{ 时};\\ 0, & x_j^k=\beta_j,\text{且 } r_j(x_N^k)<0,(j\in \bar{J})\text{ 时};\\ -r_j(x_N^k), & \text{其余情形}.(j\in \bar{J})\text{ 时},\end{cases}$$

得 $\boldsymbol{p}_N^k$.

步 3　若 $\|\boldsymbol{p}_N^k\|\leqslant\varepsilon_1$,算法终止,$\boldsymbol{x}^k$ 作为近似最优解;否则转步 4.

步 4　取 $\lambda>0$,令 $\widehat{\boldsymbol{x}}_N=\boldsymbol{x}_N^k+\lambda\,\boldsymbol{p}_N^k$,若 $\boldsymbol{\alpha}_N\leqslant\widehat{\boldsymbol{x}}_N\leqslant\boldsymbol{\beta}_N$,转步 5;否则以 $1/2\lambda$ 代替 λ,再求 $\widehat{\boldsymbol{x}}_N$,直至满足 $\boldsymbol{\alpha}_N\leqslant\widehat{\boldsymbol{x}}_N\leqslant\boldsymbol{\beta}_N$ 为止,转步 5.

步 5　用牛顿法解非线性方程组

$$H(\boldsymbol{y},\widehat{\boldsymbol{x}}_N)=\mathbf{0}.$$

具体步骤如下:

令 $\boldsymbol{y}^1=\boldsymbol{x}_B^k,j=1$,则

1) $\boldsymbol{y}^{j+1}=\boldsymbol{y}^j-(\nabla_{x_B}H(\boldsymbol{y}^j,\widehat{\boldsymbol{x}}_N))^{-1}H(\boldsymbol{y}^j,\widehat{\boldsymbol{x}}_N)$.若 $f(\boldsymbol{y}^{j+1},\widehat{\boldsymbol{x}}_N)<f(\boldsymbol{x}^k)$,$\boldsymbol{\alpha}_B\leqslant\boldsymbol{y}^{j+1}\leqslant\boldsymbol{\beta}_B$,且 $\|H(\boldsymbol{y}^{j+1},\widehat{\boldsymbol{x}}_N)\|\leqslant\varepsilon_2$,转步 6;否则转 2).

2) 若 $j=M$,则以 $1/2\lambda$ 代替 λ,令 $\widehat{\boldsymbol{x}}_N=\boldsymbol{x}_N^k+\lambda\boldsymbol{p}_N^k,\boldsymbol{y}^1=\boldsymbol{x}_B^k,j=1$,返回 1);否则 $j=j+1$,返回 1).

步 6　$\boldsymbol{x}^{k+1}=\begin{bmatrix}\boldsymbol{y}^{j+1}\\ \widehat{\boldsymbol{x}}_N\end{bmatrix}$.若 $\boldsymbol{x}^{k+1}$ 中某个基变量的值等于上界 β_j 或下界 α_j,则需将其换出基,$k=k+1$,返回步 2.

例 4　用 GRG 方法求解规划问题

$$\begin{aligned}\min\quad & x_1^2+2x_1x_2+x_2^2+12x_1-4x_2,\\ \text{s.t.}\quad & x_1^2+x_2^2\leqslant 4,\\ & 1\leqslant x_1\leqslant 3,1\leqslant x_2\leqslant 3.\end{aligned}$$

解　引入人工变量 x_3,将原问题化为如下形式:

$$\begin{aligned}\min\quad & x_1^2+2x_1x_2+x_2^2+12x_1-4x_2,\\ \text{s.t.}\quad & x_1^2+x_2^2+x_3-4=0,\\ & \begin{bmatrix}1\\1\\0\end{bmatrix}\leqslant\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}\leqslant\begin{bmatrix}3\\3\\4\end{bmatrix}.\end{aligned}$$

取初始点 $\boldsymbol{x}^0=(3/2,1,3/4)^{\mathrm{T}}$.以 x_1 为基变量,x_2,x_3 为非基变量,则通过计算得

$$\boldsymbol{r}(\boldsymbol{x}_N^0)=(-31/3,-17/3)^{\mathrm{T}},\quad \boldsymbol{p}_N^0=(31/3,17/3)^{\mathrm{T}},$$

$$\widehat{\boldsymbol{x}}_{N(\lambda)}=\boldsymbol{x}_N^0+\lambda\,\boldsymbol{p}_N^0=\begin{bmatrix}1+31\lambda/3\\3/4+17\lambda/3\end{bmatrix},$$

当 $\lambda=0.025$ 时,得 $\widehat{\boldsymbol{x}}_N=(1.258,0.892)^{\mathrm{T}}$.

解非线性方程组 $H(y,\widehat{x}_N)=0$ 得　$x^1=(1.235,1.258,0.892)^{\mathrm{T}}$. 基不变,进行下一次迭代过程.

x_1 仍为基,由计算得

$$r(x_N^1)=(-16.316,-6.877)^{\mathrm{T}},$$
$$p_N^1=(16.316,6.877)^{\mathrm{T}},$$
$$x^2=(1.01,1.421,0.961)^{\mathrm{T}}.$$

进入下一次迭代过程.

由于 x^2 的第一个分量与 λ 的第一个分量 1 比较接近,需要换基,x_2 入基,x_1 出基,类似的计算得到

$$r(x_N^2)=(16.25,-0.3)^{\mathrm{T}},$$
$$p_N^2=(0,0,3)^{\mathrm{T}},$$
$$x^3=(1.01,1.04,1.891)^{\mathrm{T}}.$$

此时目标函数值 $f(x^3)=12.16$ 已很接近最优点 $x^*=(1,1,2)^{\mathrm{T}}$ 的目标函数值 $f(x^*)=12$.

4.3　梯度投影法

对线性约束非线性规划问题

$$\min\quad f(x),$$
$$\text{s.t.}\quad Ax\geqslant b,\qquad Ex=e,$$

其中 f 为开集 $D\subset\mathbf{R}^n$ 上的连续可微函数,$A\in\mathbf{R}^{m\times n}$,$E\in\mathbf{R}^{l\times n}$,$b$ 和 e 分别为 m 维和 l 维列向量. 下面给出如下罗森(Rosen)梯度投影算法.

算法 6(罗森梯度投影法)

步 1　给定初始可行点 x^1, $k=1$.

步 2　在点 x^k 处,将 A, b 分块成

$$A=\begin{bmatrix}A_1\\A_2\end{bmatrix},\ b=\begin{bmatrix}b_1\\b_2\end{bmatrix},$$

使得

$$A_1x^k=b_1,\quad A_2x^k>b_2.$$

步 3　取 $M=\begin{bmatrix}A_1\\E\end{bmatrix}$,如果 M 是空的,则取 $P=I$(单位矩阵);否则取

$$P=I-M^{\mathrm{T}}(MM^{\mathrm{T}})^{-1}M.$$

步 4　$d^k=-p\nabla f(x^k)$,若 $d^k\neq\mathbf{0}$,转步 6;否则转步 5.

步 5　如果 M 是空的,则停止计算,x^k 作为 K-T 点;否则令 $w=(MM^{\mathrm{T}})^{-1}M\times\nabla f(x_k)=(u,v)^{\mathrm{T}}$,这里 u 的维数等于矩阵 A_1 的行数. 如果 $u\geqslant0$,则停止计算,x^k 作为 K-T 点;否则,在 u 中选择一个负的分量,如 u_{j_0},在 A_1 中去掉 u_{j_0} 所对应的行,得到新的 A_1,返回步 3.

步 6　计算求值:

$$\alpha_{max} = \begin{cases} \min\{\hat{b}_i/\hat{d}_i \mid \hat{d}_i < 0\}, & \text{当 } \hat{d} \not\geq 0 \text{ 时;} \\ \infty, & \text{当 } \hat{d} \geq 0 \text{ 时,} \end{cases}$$

其中,$\hat{b} = b_2 - A_2x^k, \hat{d} = A_2d^k$.

解规划问题

$$\begin{aligned} &\min \quad f(x^k + \alpha d^k), \\ &\text{s.t.} \quad 0 \leq \alpha \leq \alpha_{max}, \end{aligned}$$

得 α_k. 令 $x^{k+1} = x^k + \alpha_k d^k, k = k+1$,返回步 2.

例 5　用罗森梯度投影算法求解下述规划问题:

$$\begin{aligned} \min \quad & 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2, \\ \text{s.t.} \quad & -x_1 - x_2 \geq -2, \\ & -x_1 - 5x_2 \geq -5, x_j \geq 0, \quad j = 1,2. \end{aligned}$$

解　取初始可行点 $x^1 = (0,0)^T$. 由于在 x^1 处有 $I(x^1) = \{3,4\}$,故系数矩阵 A 和列向量 b 可分解为

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -1 & -1 \\ -1 & -5 \end{bmatrix},$$

$$b_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad b_2 = \begin{bmatrix} -2 \\ -5 \end{bmatrix}.$$

投影矩阵 $P = I - A_1^T(A_1A_1^T)^{-1}A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$,

$$d^1 = -P\nabla f(x^1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$w = (A_1A_1^T)^{-1}A_1\nabla f(x^1) = \begin{bmatrix} -4 \\ -6 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

修正 A_1,去掉 A_1 中 $u_2 = -6$ 对应的第二行,得新的 $A_1 = (1,0)$. 再求投影矩阵 P,

$$P = I - A_1^T(A_1A_1^T)^{-1}A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

$$d^1 = -P\nabla f(x^1) = \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$

由于

$$\hat{b} = b_2 - A_2x^1 = \begin{bmatrix} -2 \\ -5 \end{bmatrix}, \quad \hat{d} = A_2d^1 = \begin{bmatrix} -6 \\ -30 \end{bmatrix},$$

因此,

$$\alpha_{max} = \min\{-2/-6, -5/-30\} = 1/6.$$

求解规划问题

$$\min \quad 72\alpha^2 - 36\alpha,$$

$$\text{s.t.} \quad 0 \leqslant \alpha \leqslant 1/6,$$

得 $\alpha_1 = 1/6, x^2 = x^1 + \alpha_1 d^1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.转入下一个迭代过程.

对于 x^2 点,$I(x^2) = \{2,3\}$,A 和 b 可分解为

$$A_1 = \begin{bmatrix} -1 & -5 \\ 1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -1 & -1 \\ 0 & 1 \end{bmatrix},$$

$$b_1 = \begin{bmatrix} -5 \\ 0 \end{bmatrix}, \qquad b_2 = \begin{bmatrix} -2 \\ 0 \end{bmatrix},$$

从而投影矩阵

$$P = I - A_1^{\mathrm{T}}(A_1 A_1^{\mathrm{T}})^{-1} A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

这样,

$$d^2 = -P\nabla f(x^2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, W = (A_1 A_1^{\mathrm{T}})^{-1} A_1 \nabla f(x^2) = \begin{bmatrix} 2/5 \\ -28/5 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

从 A_1 中去掉 $u_2 = -28/5$ 所对应的第二行,得到新的 $A_1 = (-1, -5)$.再求投影矩阵得

$$P = I - A_1^{\mathrm{T}}(A_1 A_1^{\mathrm{T}})^{-1} A_1 = \begin{bmatrix} 25/26 & -5/26 \\ -5/26 & 1/26 \end{bmatrix},$$

从而

$$d^2 = -P\nabla f(x^2) = \begin{bmatrix} 70/13 \\ -14/13 \end{bmatrix},$$

$$\hat{b} = b_2 - A_2 x^2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix},$$

$$\hat{d} = A_2 d^2 = \begin{bmatrix} -4 \\ -1 \end{bmatrix}, \alpha_{\max} = 1/4.$$

解规划问题

$$\begin{aligned} \min \quad & 62\alpha^2 - 28\alpha - 4, \\ \text{s.t.} \quad & 0 \leqslant \alpha \leqslant 1/4, \end{aligned}$$

得 $\alpha_2 = 7/31, x^3 = x^2 + \alpha_2 d^2 = \begin{bmatrix} 35/31 \\ 24/31 \end{bmatrix}$.转入下一个迭代过程.

由于 $I(x^3) = \{2\}$,A 和 b 可分解为

$$A_1 = (-1, -5), \qquad A_2 = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$b_1 = (-5), \qquad b_2 = \begin{bmatrix} -2 \\ 0 \\ 0 \end{bmatrix},$$

从而

$$P = I - A_1^{\mathrm{T}}(A_1 A_1^{\mathrm{T}})^{-1} A_1 = 1/26 \begin{bmatrix} 25 & -5 \\ -5 & 1 \end{bmatrix},$$

$$d^3 = -P\nabla f(x^3) = \begin{bmatrix}0\\0\end{bmatrix},$$
$$w = (A_1A_1^T)^{-1}A_1\nabla f(x^3) = 32/31 > 0.$$

这样 $x^3 = \begin{bmatrix}35/31\\24/31\end{bmatrix}$ 为原规划问题的K-T点.由于本例为凸规划问题,从而也是整体极小点.

4.4　序列二次规划算法

序列二次规划方法(SQP 算法),是一类十分重要的解决带约束最优化问题(2-1)式的方法.具体算法如下:

算法 7

步 1　给定初始点 $x^1 \in \mathbf{R}^n, \sigma > 0, \delta > 0, B_1 \in \mathbf{R}^{n\times n}, \epsilon \geqslant 0$ 且满足 $\sum_{k=1}^{\infty}\epsilon_k < +\infty$ 的非负序列 $\{\epsilon_k\}, k = 1$.

步 2　解二次规划子问题

$$\min \quad \nabla f(x^k)^T d + \frac{1}{2}d^T B_k d, \tag{4-5}$$
$$\text{s.t.} \quad \nabla c_i(x^k)^T d + c_i(x^k) = 0, i \in E,$$
$$\nabla c_i(x^k)^T d + c_i(x^k) \geqslant 0, i \in I,$$

得 d^k.若 $\|d^k\| \leqslant \epsilon$,算法终止;否则求 α_k 满足:

$$P(x^k + \alpha_k d^k, \sigma) \leqslant \min\{P(x^k + \alpha d^k, \sigma) \mid 0 \leqslant \alpha \leqslant \delta\} + \epsilon_k,$$

其中

$$P(x,\sigma) = f(x) + \sigma\left[\sum_{i\in E}|c_i(x)| + \sum_{i\in I}|c_i(x)_-|\right].$$

步 3　$x^{k+1} = x^k + \alpha_k d^k$,计算 B_{k+1} 如下:

$$s^k = x^{k+1} - x^k,$$
$$y^k = \nabla f(x^{k+1}) - \nabla f(x^k) - \sum_{i\in E\cup I}\lambda_i^k[\nabla c_i(x^{k+1}) - \nabla c_i(x^k)],$$
$$B_{k+1} = B_k - \frac{B_k s^k {y^k}^T + y^k {s^k}^T B_k}{{s^k}^T y^k} + \left(1 + \frac{{s^k}^T B_k s^k}{{s^k}^T y^k}\right)\frac{y^k {y^k}^T}{{s^k}^T y^k},$$

$k = k + 1$,返回步 2.

其中 λ^k 是二次规划子问题(4-5)的拉格朗日乘子.

算法分析　在一定条件下,SQP算法具有全局收敛性和超线性的收敛速度.但上述方法有时会出现马洛托斯效应,即由于引入 L_1 精确罚函数 $p(x,\sigma)$,而可能破坏超线性收敛性.

参 考 文 献

1　巴扎拉 M S,希蒂 C M著.非线性规划.贵阳:贵州人民出版社,1985.
2　邓乃扬著.无约束最优化计算方法.北京:科学出版社,1988.
3　陈宝林著.最优化理论与算法.北京:清华大学出版社,1989.
4　席少霖,赵凤治著.最优化计算方法.上海:上海科技出版社,1983.
5　赵瑞安,吴方著.非线性最优化理论和方法.杭州:浙江科学技术出版社,1992.
6　袁亚湘,孙文瑜著.最优化理论与方法.北京:科学出版社,1997.
7　陈开明编著.非线性规划.上海:复旦大学出版社,1991.
8　袁亚湘著.非线性规划数值方法.上海:上海科学技术出版社,1992.

·经济数学卷·

第8篇

不可微优化

编　者　高　岩
审校者　夏尊铨

目　录

引　言

不可微优化又称非光滑优化，它是最优化理论与方法中的一个重要分支．所谓不可微优化，是指目标函数或约束函数中至少有一个不是连续可微(光滑)的非线性规划问题．由于缺少连续可微(光滑)性质，经典的基于梯度概念的非线性规划理论和算法不再适用于不可微优化问题．对梯度(微分)概念进行推广，在推广的梯度理论基础上建立相应的最优性理论和数值计算方法，正是不可微优化研究工作的主要目的．

在众多的广义梯度理论中，关于局部利普希茨(Lipschitz)函数的克拉克(D.A.Clarke)广义梯度，是目前研究不可微优化的主要工具之一．本篇以克拉克广义梯度及其微分学为理论基础，对局部利普希茨函数的优化理论和方法进行较详细的介绍．对其他几种广义梯度(微分)，例如：拟微分、迪尼(Dini)导数、贝诺特(Penot)微分及其最优性理论只作扼要介绍．

1　基本概念和理论基础

1.1　不可微函数和不可微优化举例

(1) 设 $x_1,\cdots,x_N\in\mathbf{R}^n$ 为 N 组实验数据，要建立一个线性模型，即求出一个超平面

$$H=\{x\in\mathbf{R}^n\mid\langle a,x\rangle=b\},$$

其中，$a\in\mathbf{R}^n$，$b\in\mathbf{R}^1$，使得 $x_1,\cdots,x_N$ 尽可能接近 H，这就引出一个不可微优化问题

$$\begin{cases}\min\sum_{k=1}^{N}\left|\sum_{i=1}^{n}a_ix_k^i-b\right|;\\ a\in\mathbf{R}^n,b\in\mathbf{R}^1,\end{cases}\tag{1-1}$$

其中，a_i，x_k^i 分别为 a 和 x_k 的第 i 个分量．

但是通常为回避不可微困难，考虑下述问题：

$$\begin{cases}\min\sum_{k=1}^{N}\left(\sum_{i=1}^{n}a_ix_k^i-b\right)^2;\\ a\in\mathbf{R}^n,b\in\mathbf{R}^1.\end{cases}\tag{1-2}$$

这就是通常所说的最小二乘问题．一般来讲，(1-1) 式较(1-2) 式要合理，下面举例说明．取 $n=2$，$x_1=(1,1)$，$\cdots$，$x_{N-1}=(N-1,N-1)$，$x_N=(N,0)$．前$(N-1)$个点都落在直线

$$L = \{(x^1, x^2) \in \mathbf{R}^2 \mid x^1 = x^2\}$$

上，第 N 个点不在此直线上. 不难看出，当 $N \geqslant 3$ 时，由(1-1)式得到的解正是直线 L，而用最小二乘法(1-2)式则得到另外一条直线.

(2) 设 C 为 $\mathbf{R}^n$ 中非空闭集，点 $x \in \mathbf{R}^n$ 到 C 的距离定义为

$$d_C(x) = \min_{y \in C} \| x - y \| .$$

故函数 $d_C(x)$ 是不可微的.

(3) 考虑约束优化问题

$$\begin{cases} \min f(x), \\ \text{s.t. } g(x) \leqslant 0, \end{cases} \tag{1-3}$$

其中，$f(x)$，$g(x)$ 为在 $\mathbf{R}^n$ 上的连续可微函数. 利用罚函数，可将求解约束优化问题(1-3)转化为求解无约束优化问题

$$\min_{x \in \mathbf{R}^n} f(x) + \mu \max\{0, g(x)\}, \tag{1-4}$$

其中 $\mu > 0$. 显然，(1-4)式中目标函数是不可微函数.

(4) 考虑非线性互补问题

$$h(x) \geqslant 0, \quad f(x) \geqslant 0, \quad h(x)^{\mathrm{T}} f(x) = 0, \tag{1-5}$$

其中 $h(x) = (h_1(x), \cdots, h_n(x))^{\mathrm{T}}$，$f(x) = (f_1(x), \cdots, f_n(x))^{\mathrm{T}}$，$h_i(x)$，$f_i(x)$ 均为 $\mathbf{R}^n$ 上的连续可微函数. 求解问题(1-5)可转化为求解非光滑方程组

$$\min\{h_i(x), f_i(x)\} = 0, \quad i = 1, 2, \cdots, n. \tag{1-6}$$

求解(1-6)式等价于求解优化问题

$$\min_{x \in \mathbf{R}^n} \sum_{i=1}^{n} (\min\{h_i(x), f_i(x)\})^2. \tag{1-7}$$

(1-7)式是一个不可微优化问题. 不过近年来人们比较倾向于通过求解非光滑方程组(1-6)来解非线性互补问题(1-5).

1.2　集值映射

不可微优化的理论基础是各种推广的梯度及其相应的微分学理论. 广义梯度一般说来已不再是单值映射，通常是集值映射.

F 称为 $\mathbf{R}^n$ 到 $\mathbf{R}^m$ 上的**集值映射**是指：给定 $x \in \mathbf{R}^n$，$F(x)$ 为 $\mathbf{R}^m$ 中一子集，集合

$$\mathrm{dom} F = \{x \in \mathbf{R}^n \mid F(x) \neq \varnothing\}$$

称为 F 的有效域. $\mathrm{dom} F \neq \varnothing$ 的集值映射称为真集值映射，$\mathbf{R}^{m+n}$ 中子集

$$\mathrm{graph} F = \{(x, y) \in \mathbf{R}^{m+n} \mid y \in F(x)\}$$

称为 F 的图像.

集值映射 F 在 x 点称为**上半连续的**，如果对任意 $\epsilon > 0$，存在 $\delta > 0$，使得

$$F(x') \subset F(x) + B(0, \epsilon), \forall x' \in B(x, \delta),$$

其中 $B(x, \delta) = \{x' \in \mathbf{R}^n \mid \| x' - x \| \leqslant \delta\}$.

若 F 在 Ω 中每一点都是上半连续的，则称 F 在 Ω 中上半连续.

集值映射 F 在 x 点称为**下半连续的**，如果对任意 $\epsilon > 0$，存在 $\delta > 0$，使得

$$F(x)+B(0,\epsilon)\subset F(x'),\forall x'\in B(x,\delta).$$

若 F 在 Ω 中每一点都是下半连续的，则称 F 在 Ω 中下半连续.

例1 设 F 为 $\mathbf{R}^1$ 到 $\mathbf{R}^1$ 上的集值映射，定义如下：

$$F(x)=\begin{cases}\{0\}, & \text{当 } x\neq 0;\\ [-1,1], & \text{当 } x=0.\end{cases}$$

F 在 $x=0$ 处上半连续.

例2 设 F 为 $\mathbf{R}^1$ 到 $\mathbf{R}^1$ 上的集值映射，定义如下：

$$F(x)=\begin{cases}[-1,1], & \text{当 } x\neq 0;\\ \{0\}, & \text{当 } x=0.\end{cases}$$

F 在 $x=0$ 处下半连续.

半连续性是集值映射中的一个重要概念.在不可微优化中，经常遇到的是上半连续性，这主要是由于最常用到的克拉克广义梯度是上半连续的，而众多的不可微优化算法的收敛性研究都利用了这一性质.此外，还有一个重要概念就是集值映射的闭性.

集值映射在 x 点称为闭的，若对任意序列 $\{x_k\}_1^\infty$ 满足

$$x_k\to x,\ y_k\to y,\ y_k\in F(x_k),$$

必有 $y\in F(x)$.如果 F 在 Ω 上每一点都是闭的，则称 F 为 Ω 上闭映射.

通常人们对集值映射的上半连续性和闭性是不加区别的，视二者为一个概念.事实上，若集值映射 F 在 x 处局部一致有界，即存在 x 的一个邻域 $N(x)$，在此邻域内 $\bigcup_{x'\in N(x)}F(x')$ 有界，则 F 在 x 处上半连续等价于 F 在 x 处是闭的.

设 F 为 $\mathbf{R}^n$ 到 $\mathbf{R}^n$ 上的凸紧集值映射，即对任意 $x\in\mathbf{R}^n$，$F(x)$ 为 $\mathbf{R}^n$ 中一致有界的凸紧集，定义**支撑函数**

$$\delta^*(l\mid F(x))=\max_{y\in F(x)}\langle y,l\rangle,\forall l\in\mathbf{R}^n.$$

对集值映射，有下述定理：

定理1 上述集值映射 F 在 $x\in\mathbf{R}^n$ 处上半连续的充要条件是，对每一 $l\in\mathbf{R}^n$，$\delta^*(l\mid F(\cdot))$ 在 x 处上半连续.

证明 必要性 设 F 在 x 处上半连续，取固定的 $l\in\mathbf{R}^n$，对任意满足 $x_k\to x$ 的点列 $(x_k)_1^\infty$，由于 $F(x_k)$ 是紧的，则存在 $y_k\in F(x_k)$ 使得

$$\delta^*(l\mid F(x_k))=\max_{y\in F(x_k)}\langle y,l\rangle=\langle y_k,l\rangle.$$

点列 $\{x_k\}_1^\infty$ 有界，由 F 的一致有界性，点列 $\{y_k\}_1^\infty$ 也是有界的.设 $\{k_i\}$ 为 $\{k\}$ 中满足 $\delta^*(l\mid F(x_{k_i}))\to\overline{\lim_{k\to\infty}}\delta^*(l\mid F(x_k))$ 的子列，又 $\{y_{k_i}\}$ 有界，必有收敛子列，不妨假设 $\lim_{k_i\to\infty}y_{k_i}=y$.由 F 的上半连续性有 $y\in F(x)$，于是

$$\begin{aligned}\overline{\lim_{k\to\infty}}\delta^*(l\mid F(x_k))&=\lim_{k_i\to\infty}\delta^*(l\mid F(x_{k_i}))=\lim_{k_i\to\infty}\langle y_{k_i},l\rangle\\&=\langle y,l\rangle\leqslant\max_{y'\in F(x)}\langle y',l\rangle=\delta^*(l\mid F(x)),\end{aligned}$$

这就是说 $\delta^*(l\mid F(\cdot))$ 上半连续.

充分性 设 $\delta^*(l \mid F(\cdot))$ 在 x 处上半连续.假设 F 在 x 处不是上半连续的,则存在点列 $\{x_k\}_1^\infty$, $\{y_k\}_1^\infty$,使得 $x_k \to x, y_k \to y, y_k \in F(x_k), y \notin F(x)$.由于 $F(x)$ 为凸紧集,根据凸分析中分离定理,存在 $l \in \mathbf{R}^n$ 和 $\epsilon > 0$,使得

$$\langle y, l\rangle \geqslant \max_{y' \in F(x)} \langle y', l\rangle + \epsilon = \delta^*(l \mid F(x)) + \epsilon.$$

对充分大的 k,有

$$\delta^*(l \mid F(x_k)) \geqslant \langle y_k, l\rangle > \langle y, l\rangle - \frac{\epsilon}{2} > \delta^*(l \mid F(x)) + \frac{\epsilon}{2},$$

这与 $\delta^*(l \mid F(\cdot))$ 的上半连续性矛盾.充分性得证.

对于集值映射还有下述不动点定理:

定理2 设 F 为凸紧集 $\Omega \subset \mathbf{R}^n$ 上的集值映射,即对任意 $x \in \Omega$, $F(x)$ 为 Ω 中凸紧集.如果 F 在 Ω 中上半连续,则存在 $x_0 \in \Omega$,使得 $x_0 \in F(x_0)$.

2 广义梯度理论

2.1 局部利普希茨函数广义梯度

设 $f(x)$ 为定义于 $\Omega \subset \mathbf{R}^n$ 上的实函数,若对任意 $x \in \Omega$,存在常数 $\delta_x, L_x > 0$,使得

$$|f(x_1) - f(x_2)| \leqslant L_x \|x_1 - x_2\|,$$

$$\forall x_1, x_2 \in B(x, \delta_x) = \{x' \in \mathbf{R}^n \mid \|x' - x\| < \delta_x\},$$

则称 $f(x)$ 为 Ω 上的**局部利普希茨函数**.若在上式中常数 L_x 不依赖于 x,则称 $f(x)$ 为 Ω 上的(整体)利普希茨函数.

命题1 连续可微函数是局部利普希茨函数.

命题2 设 $f_1(x), f_2(x)$ 是 $\Omega \subset \mathbf{R}^n$ 上局部利普希茨函数,那么对任意实数 λ_1, λ_2,有 $\lambda_1 f(x) + \lambda_2 f(x)$ 是 Ω 上局部利普希茨函数.

命题2说明局部利普希茨函数全体构成一个线性空间.

命题3 设 $f_1(x), f_2(x)$ 是 Ω 上局部利普希茨函数,则

$$g(x) = \max\{f_1(x), f_2(x)\},$$
$$h(x) = \min\{f_1(x), f_2(x)\}$$

也是 Ω 上的局部利普希茨函数.

证明 由

$$|g(x_1) - g(x_2)|, \quad |h(x_1) - h(x_2)|$$
$$\leqslant \max\{|f_1(x_1) - f_1(x_2)|, |f_2(x_1) - f_2(x_2)|\}$$

易见, $g(x), h(x)$ 为局部利普希茨函数.

命题4 设 $f(x)$ 是定义于凸开集 $\Omega \subset \mathbf{R}^n$ 上的凸函数,那么 $f(x)$ 在 Ω 上是局部利普希茨函数.

命题5 设 C 为 $\mathbf{R}^n$ 中非空闭集，$d_C(x)$ 表示点 x 到 C 的距离，那么函数 $d_C(x)$ 是 $\mathbf{R}^n$ 上的(整体)利普希茨函数.

证明 对任意 $x_1, x_2 \in \mathbf{R}^n$，由于 C 为非空闭集，必存在 $y_1 \in C$，使得

$$d_C(x_1) = \| x_1 - y_1 \|.$$

于是

$$\begin{aligned} d_C(x_2) &\leqslant \| x_2 - y_1 \| \leqslant \| x_1 - x_2 \| + \| x_1 - y_1 \| \\ &= \| x_1 - x_2 \| + d_C(x_1), \end{aligned}$$

$$d_C(x_2) - d_C(x_1) \leqslant \| x_1 - x_2 \|. \tag{2-1}$$

同理有

$$d_C(x_1) - d_C(x_2) \leqslant \| x_1 - x_2 \|. \tag{2-2}$$

综合(2-1)式和(2-2)式，有

$$| d_C(x_1) - d_C(x_2) | \leqslant \| x_1 - x_2 \|.$$

也就是说，$d_C(x)$ 在 $\mathbf{R}^n$ 上是利普希茨函数.

以上几个命题说明，局部利普希茨函数具有相当的广泛性.就不可微优化而言，目前研究最多的非光滑函数就是局部利普希茨函数.针对局部利普希茨函数，克拉克提出了广义梯度概念，作为光滑函数梯度和凸函数次微分的推广，广义梯度及其微分学是目前非光滑函数分析和优化中最能被人们接受的一种理论.

设 $f(x)$ 为开集 $\Omega \subset \mathbf{R}^n$ 上的局部利普希茨函数，$f(x)$ 的广义方向导数定义如下：

$$f^\circ(x;h) = \varlimsup_{\substack{y \to x \\ t \to 0^+}} \frac{f(y+th) - f(y)}{t}, \quad \forall h \in \mathbf{R}^n, \tag{2-3}$$

$f^\circ(x;h)$ 通常称为**克拉克广义方向导数**.

对于局部利普希茨函数，上述极限总是存在的，换句话说，克拉克广义方向导数总是存在的.

定理1 设 $f(x)$ 为开集 $\Omega \subset \mathbf{R}^n$ 上的局部利普希茨函数，则有下述结论：

1° $f^\circ(x;\cdot)$ 有限，正齐次，次可加，且满足 $|f^\circ(x;h)| \leqslant L_x \|h\|$，其中 L_x 为 f 在 x 点附近的利普希茨常数.

2° $f^\circ(x;h)$ 作为 (x,h) 的函数是上半连续的；作为 h 的函数是局部利普希茨函数.

3° $f^\circ(x;-h) = (-f)^\circ(x;h), \forall h \in \mathbf{R}^n$.

证明 由 $f(x)$ 的局部利普希茨性质，当 y 充分接近 x 点，t 充分小时，有

$$\frac{|f(y+th) - f(y)|}{t} \leqslant L_x \|h\|.$$

再由 $f^\circ(x;h)$ 定义易见，$|f^\circ(x;h)| \leqslant L_x \|h\|$，即 $f^\circ(x;\cdot)$ 有界.关系式 $f^\circ(x;\lambda h) = \lambda f^\circ(x;h), \forall \lambda > 0$，由广义方向导数定义是显而易见的.通过计算得

$$f^\circ(x;h_1+h_2) = \varlimsup_{\substack{y \to x \\ t \to 0^+}} \frac{f(y+th_1+th_2) - f(y)}{t}$$

$$\leqslant \overline{\lim_{\substack{y \to x \\ t \to 0^+}}} \frac{f(y + th_1 + th_2) - f(y + th_2)}{t} + \overline{\lim_{\substack{y \to x \\ t \to 0^+}}} \frac{f(y + th_2) - f(y)}{t}$$

$$= f^{\circ}(x;h_1) + f^{\circ}(x;h_2),$$

$f^{\circ}(x;\cdot)$ 是正齐次的，1° 得证.

设 $x_i \to x, h_i \to h(i \to \infty)$. 对每一个 i，由广义方向导数定义，存在 y_i 和 t_i 使得

$$\| y_i - x_i \| + t_i < \frac{1}{i},$$

$$f^{\circ}(x_i;h_i) - \frac{1}{i} \leqslant \frac{f(y_i + t_i h_i) - f(y_i)}{t_i}$$
$$= \frac{f(y_i + t_i h) - f(y_i)}{t_i} + \frac{f(y_i + t_i h_i) - f(y_i + t_i h)}{t_i}.$$

注意到，$L_x \| h_i - h \|$ 可作为上式后一项的界，取极限得

$$\overline{\lim_{i \to \infty}} f^{\circ}(x_i;h_i) \leqslant f^{\circ}(x;h),$$

即 $f^{\circ}(x;h)$ 关于 (x,h) 是上半连续的. 另一方面，对充分接近 x 点和 0 的 y 和 t，有

$$f(y + th_1) - f(y) \leqslant f(y + th_2) - f(y) + L_x \| h_1 - h_2 \| t,$$

对上式取极限得

$$f^{\circ}(x;h_1) \leqslant f^{\circ}(x;h_2) + L_x \| h_1 - h_2 \|.$$

上式中 h_1 和 h_2 是对称的，易见 $f^{\circ}(x;\cdot)$ 是利普希茨函数，2° 得证.

因为

$$f^{\circ}(x;-h) = \overline{\lim_{\substack{y \to x \\ t \to 0^+}}} \frac{f(y - th) - f(y)}{t}$$
$$= \overline{\lim_{\substack{y \to x \\ t \to 0^+}}} \frac{(-f)(y) - (-f)(y - th)}{t}.$$

令 $u = y - th$，则 $y = u + th$，而当 $y \to x$ 时，有 $u \to x$，所以

$$f^{\circ}(x;-h) = \overline{\lim_{\substack{u \to x \\ t \to 0^+}}} \frac{(-f)(u + th) - (-f)(u)}{t} = (-f)^{\circ}(x;h),$$

3° 式得证.

一般说来，局部利普希茨函数的普通方向导数

$$f'(x;h) = \lim_{t \to 0^+} \frac{f(x + th) - f(x)}{t}, \quad h \in \mathbf{R}^n$$

不一定存在，即使存在，它与克拉克广义方向导数也未必相等.

例 1　设　$f(x)=\begin{cases}x^2\sin\dfrac{1}{x}, & 当\ x\neq 0;\\ 0, & 当\ x=0.\end{cases}$

不难验证，$f'(0;\pm 1)=0$，而 $f^\circ(0;\pm 1)=1$.

例 2　设　$f(x)=|x|$，容易验证

$$f^\circ(0;1)=f'(0;1)=1, f^\circ(0;-1)\neq f'(0;-1)=-1.$$

对于局部利普希茨函数 $f(\boldsymbol{x})$，若其方向导数 $f'(\boldsymbol{x};\boldsymbol{h}),\forall\boldsymbol{h}\in\mathbf{R}^n$ 存在，且 $f^\circ(\boldsymbol{x};\boldsymbol{h})=f'(\boldsymbol{x};\boldsymbol{h}),\forall\boldsymbol{h}\in\mathbf{R}^n$，即克拉克广义方向导数与普通方向导数相等，此时称 $f(\boldsymbol{x})$ 是**正则的**.

如果 $f_1(\boldsymbol{x}),f_2(\boldsymbol{x})$ 是正则的，对任意常数 $\lambda_1,\lambda_2\geqslant 0,\lambda_1f_1(\boldsymbol{x})+\lambda_2f_2(\boldsymbol{x})$ 也是正则的. 连续可微函数是正则的，凸函数是正则的，凹函数不是正则的(除非它是可微的).

借助广义方向导数，定义局部利普希茨函数广义梯度如下：

$$\partial f(\boldsymbol{x})=\{\boldsymbol{\xi}\in\mathbf{R}^n\mid\langle\boldsymbol{\xi},\boldsymbol{h}\rangle\leqslant f^\circ(\boldsymbol{x};\boldsymbol{h}),\forall\boldsymbol{h}\in\mathbf{R}^n\},\qquad(2\text{-}4)$$

$\partial f(\boldsymbol{x})$ 通常称为**克拉克广义梯度**. 广义梯度 $\partial f(\boldsymbol{x})$ 是 $\mathbf{R}^n$ 中的凸紧集，它与广义方向导数有下述关系

$$f^\circ(\boldsymbol{x};\boldsymbol{h})=\max\{\langle\boldsymbol{\xi},\boldsymbol{h}\rangle\mid\boldsymbol{\xi}\in\partial f(\boldsymbol{x})\},\forall\boldsymbol{h}\in\mathbf{R}^n.\qquad(2\text{-}5)$$

当 $f(\boldsymbol{x})$ 是连续可微函数时，克拉克广义梯度为单点集，且有 $\partial f(\boldsymbol{x})=\{\nabla f(\boldsymbol{x})\}$；当 $f(\boldsymbol{x})$ 是凸函数时，克拉克广义梯度 $\partial f(\boldsymbol{x})$ 就是**凸函数的次微分**.

定理 2　设 $f(\boldsymbol{x})$ 是开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，则集值映射 $\boldsymbol{x}\to\partial f(\boldsymbol{x})$ 是上半连续的.

证明　利用第一章 1.2 节定理 1 及本章 2.1 节定理 1 中的 1° 及(2-5)式，立刻得到本定理.

克拉克广义梯度的上半连续性质在非光滑分析和不可微优化理论与算法研究中具有重要意义.

推论 1　设 $f(\boldsymbol{x})$ 为定义在开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，若其克拉克广义梯度 $\partial f(\boldsymbol{x})$ 在 Ω 上恒为单点集，则 $f(\boldsymbol{x})$ 在 Ω 上连续可微，且有 $\partial f(\boldsymbol{x})=\{\nabla f(\boldsymbol{x})\}$.

关于有限维空间上局部利普希茨函数可微性有下述定理.

定理 3　设 $f(\boldsymbol{x})$ 是开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，那么 $f(\boldsymbol{x})$ 在 Ω 上几乎处处可微，即 $f(\boldsymbol{x})$ 在 Ω 上除掉一个测度为零的集合外是可微的.

定理 4　设 $f(\boldsymbol{x})$ 是开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，则有

$$\partial f(\boldsymbol{x})=\mathrm{co}\{\lim_{\boldsymbol{x}_i\to\boldsymbol{x}}\nabla f(\boldsymbol{x}_i)\mid f(\boldsymbol{x})\ 在\ \boldsymbol{x}=\boldsymbol{x}_i\ 处可微\}.\qquad(2\text{-}6)$$

定理 4 建立了广义梯度与其邻域内可微点梯度之间的关系，甚至(2-6)式可作为广义梯度的一个等价定义.

例 3　设 $f(x)=|x|,x\in\mathbf{R}^1$. 显然，$f(x)$ 在 $x=0$ 处是不可微的，当 $x>0$ 时，$\nabla f(x)=1$；当 $x<0$ 时，$\nabla f(x)=-1$. 于是

$$\begin{aligned}\partial f(0)&=\mathrm{co}\{\lim_{x_i\to 0}\nabla f(x_i)\mid f(x)\ 在\ x=x_i\ 处可微\}\\&=\mathrm{co}\{1,-1\}=[-1,1].\end{aligned}$$

2.2　广义梯度运算法则

命题6　设 $f(x),g(x)$ 为开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，那么有

1° $\partial(\lambda f)(x)=\lambda\partial f(x),\forall\lambda\in\mathbf{R}^1$；

2° $\partial(f(x)+g(x))\subset\partial f(x)+\partial g(x)$；

3° 当 $f(x)$ 在 x 处连续可微或 $f(x)$ 和 $g(x)$ 都是凸函数时，2° 转化为等式.

证明　利用定理4中(2-6)式易见1°成立.利用

$$(f(x;h)+g(x;h))^\circ\leqslant f^\circ(x;h)+g^\circ(x;h),\forall h\in\mathbf{R}^n$$

和凸分析中凸紧集与支撑函数的一一对应关系，可得2°.当 $f(x)$ 在 x 处可微时，有

$$(f(x;h)+g(x;h))^\circ=\langle\nabla f(x),h\rangle+g^\circ(x;h);$$

而当 $f(x),g(x)$ 都是凸函数时，有

$$(f(x;h)+g(x;h))^\circ=f'(x;h)+g'(x;h).$$

再由凸紧集与支撑函数一一对应关系可知3°成立.

命题7　设 $f(x),g(x)$ 是开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，那么

1° $f(x),g(x)$ 在 Ω 上是局部利普希茨函数，且有

$$\partial(f(x)g(x))\subset f(x)\partial g(x)+g(x)\partial f(x).$$

2° 当 $g(x)\neq0$ 时，$f(x)/g(x)$ 也是局部利普希茨函数，且有

$$\partial\left(\frac{f(x)}{g(x)}\right)\subset\frac{g(x)\partial f(x)-f(x)\partial g(x)}{g^2(x)}.$$

命题8　设 $f(x)$ 在 x 的邻域内是局部利普希茨函数，$h(y)$ 在 $y=f(x)$ 的邻域内是连续可微的，那么

$$\partial(h(x)\circ f(x))=\nabla h(f(x))\partial f(x).$$

命题9　设 $g=(g_1,g_2,\cdots,g_m)^{\mathrm{T}}$，其中 $g_i,i=1,2,\cdots,m$ 是 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，$f(y)$ 是 $g(\Omega)=\{(y_1,y_2,\cdots,y_m)\mid y_i=g_i(x),x\in\Omega\}$ 上的局部利普希茨函数，则 $f\circ g(x)$ 是 Ω 上的局部利普希茨函数，且有

$$\xi_i\partial(f\circ g(x))\subset\overline{\mathrm{co}}\left\{\sum_{i=1}^{m}\alpha_i\xi_i\mid\xi_i\in\partial g_i(x),\alpha\in\partial f(g(x)),i=1,2,\cdots,m\right\},\tag{2-7}$$

其中 $\overline{\mathrm{co}}$ 表示闭凸包，α_i 为 α 的第 i 个分量.

命题10　设 $f_i(x),i=1,2,\cdots,m$ 在 x 的邻域内是局部利普希茨函数，令

$$f(x)=\max_{1\leqslant i\leqslant m}f_i(x),$$

$$I(x)=\{i\mid f_i(x)=f(x),1\leqslant i\leqslant m\},$$

那么 $f(x)$ 在 x 的邻域内也是局部利普希茨函数，且有

$$\partial f(x)\subset\mathrm{co}\{\partial f_i(x)\mid i\in I(x)\}.\tag{2-8}$$

如果 $f_i(x)$ 都是连续可微函数或都是凸的，(2-8)式转化为等式，即

$$\partial f(x)=\mathrm{co}\{\partial f_i(x)\mid i\in I(x)\}.\tag{2-9}$$

利用(2-9)式，很容易得到 $f(x)=|x|$ 在 $x=0$ 处的克拉克广义梯度.将 $f(x)$ 表

示为 $f(x)=\max\{x,-x\}$，于是有

$$\partial f(0)=\mathrm{co}\{\partial x\mid_{x=0},\partial(-x)\mid_{x=0}\}=\mathrm{co}\{1,-1\}=[-1,1].$$

更一般地，有下述定理.

定理 5　设 $f(x)=\sup\limits_{t\in T}f_t(x)$，其中 $f_t(x),t\in T$ 是开集 $\Omega\subset\mathbf{R}^n$ 上的局部利普希茨函数，T 为任意指标集，且存在常数 $L>0$，使得

$$|f_t(x_1)-f_t(x_2)|\leqslant L\|x_1-x_2\|,\forall x_1,x_2\in\Omega,t\in T,$$

那么 $\partial f(x)\subset\mathrm{co}\{\lim\limits_{x_i\to x}\nabla f_{t_i}(x_i)\mid t_i\in T,f_{t_i}(x)$ 在 $x=x_i$ 处可微，$x_i\to x,f_{t_i}(x_i)\to f(x)\}$.

定理 6　设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，如果 $f(x)$ 在 $x=x^*$ 处达到局部极值，那么有 $\mathbf{0}\in\partial f(x^*)$.

证明　由于 $\partial(-f)(x)=-\partial f(x)$，所以只需考虑极小点情形. 设 x^* 为 $f(x)$ 的极小点，于是存在 $\epsilon>0$，使得 $f(x^*+th)-f(x^*)\geqslant 0,\forall 0<t<\epsilon,h\in\mathbf{R}^n$.

进一步有

$$\begin{aligned}f^\circ(x^*;h)&=\overline{\lim_{\substack{y\to x^*\\t\to0^+}}}\frac{f(y+th)-f(y)}{t}\\&\geqslant\overline{\lim_{t\to0^+}}\frac{f(x^*+th)-f(x^*)}{t}\\&\geqslant0,\forall h\in\mathbf{R}^n.\end{aligned}\tag{2-10}$$

广义方向导数 $f^\circ(x^*;\cdot)$ 是广义梯度 $\partial f(x^*)$ 的支撑函数，根据凸紧集与其支撑函数一一对应关系，有 $\mathbf{0}\in\partial f(x^*)$.

例 4　设 $f(x)=|\sin x|,x\in\mathbf{R}^1$. 易见

$$\partial f(x)=\begin{cases}\{|\cos x|\}, & \text{当 }\sin x\neq0;\\ [-1,1], & \text{当 }\sin x=0.\end{cases}$$

$x=\frac{\pi}{a}+k\pi$（k 为整数）是 $f(x)$ 的极大值点，$\partial f(\frac{\pi}{2}+k\pi)=\{\cos(\frac{\pi}{2}+k\pi)\}=\{0\}$；$x=k\pi$（$k$ 为整数）是 $f(x)$ 的极小值点，$\partial f(k\pi)=[-1,1]$.

利用定理 6 可导出下述**中值定理**.

定理 7　设 $x,y\in\mathbf{R}^n$，f 是在包含线段 $[x,y]$ 的某一开集上的局部利普希茨函数，则存在线段 (x,y) 中一点 ξ，使得

$$f(y)-f(x)\in\langle\partial f(\xi),y-x\rangle.$$

证明　定义函数

$$g(t)=f(x+t(y-x)),t\in[0,1],$$

易证，$g(t)$ 是局部利普希茨函数. 首先证明不等式

$$\partial g(t)\subset\langle\partial f(x+t(y-x)),y-x\rangle.\tag{2-11}$$

记 $x_t=x+t(y-x)$，由于 $\partial g(t)$ 和 $\langle\partial f(x+t(y-x)),y-x\rangle$ 均为 $\mathbf{R}^1$ 中闭区间，欲证(2-11)式，只需证明

$$\delta^*(h\mid\partial g(t))\leqslant\delta^*(h\mid\langle\partial f(x_t),y-x\rangle),h=\pm1.\tag{2-12}$$

由于 $\delta^*(h \mid \partial g(t)) = g^\circ(t;h)$，于是

$$\begin{aligned}\delta^*(h \mid \partial g(t)) &= \overline{\lim_{\substack{s\to t\\ \lambda\to 0^+}}} \frac{g(s+\lambda h)-g(s)}{\lambda}\\ &= \overline{\lim_{\substack{s\to t\\ \lambda\to 0^+}}} \frac{f(x+(s+\lambda h)(y-x))-f(x+s(y-x))}{\lambda}\\ &= \overline{\lim_{\substack{y'\to x\\ \lambda\to 0^+}}} \frac{f(y'+\lambda h(y-x))-f(y')}{\lambda}\\ &= f^\circ(x_t;h(y-x))\\ &= \delta^*(h \mid \langle \partial f(x_t), y-x\rangle),\ h=\pm 1,\end{aligned} \tag{2-13}$$

(2-12)式成立，于是(2-11)式成立. 得证.

定义函数 $G(t) = g(t) + t(f(x) - f(y))$，$G(t)$ 为闭区间$[0,1]$上的连续函数，且 $G(0) = G(1) = f(x)$，于是存在 $t_0 \in (0,1)$，使得 $G(t)$ 在 $t = t_0$ 处达到极值. 根据定理6，$0 \in \partial G(t_0)$. 注意到 $\partial G(t) \subset \partial g(t) + \{f(x) - f(y)\}$，再由(2-11)式得

$$0 \in \langle \partial f(x + t_0(y-x)), y-x\rangle + \{f(x) - f(y)\}.$$

令 $\xi = x + t_0(y - x)$，即得定理结论.

2.3　ε 广义梯度和广义雅可比

设 $f(x)$ 为 $\mathbf{R}^n$ 上的局部利普希茨函数，给定 $\varepsilon \geqslant 0$，称集合

$$\partial_\varepsilon^G f(x) = \mathrm{co}\{\partial f(y) \mid y \in B(x,\varepsilon)\}$$

为 $f(x)$ 的 **ε 广义梯度**，其中 $B(x,\varepsilon) = \{y \in \mathbf{R}^n \mid \|y - x\| \leqslant \varepsilon\}$.

定理8　设 $f(x)$ 是开集 $\Omega \subset \mathbf{R}^n$ 上的局部利普希茨函数，利普希茨常数为 L_x，有下述结论：

1° $\partial_0^G f(x) = \partial f(x)$；

2° $\partial_{\varepsilon_1}^G f(x) \subset \partial_{\varepsilon_2}^G f(x)$，$\varepsilon_1 \leqslant \varepsilon_2$；

3° $\partial_\varepsilon^G f(x)$ 为非空凸紧集，且有 $\|\xi\| \leqslant L_x$，$\forall \xi \in \partial_\varepsilon^G f(x)$；

4° 集值映射 $x \to \partial_\varepsilon^G f(x)$ 是上半连续的.

命题11　设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，那么

$$\partial_\varepsilon^G f(x) = \mathrm{co}\{\lim_{y_i\to y} \nabla f(y_i) \mid f(x) \text{ 在 } x = y_i \text{ 处可微}, y \in B(x,\varepsilon)\}.$$

命题12　设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，那么

$$\partial f(y) \subset \partial_\varepsilon^G f(x),\ \forall y \in B(x,\varepsilon).$$

设 $F: \mathbf{R}^n \to \mathbf{R}^m$ 为向量局部利普希茨函数，记 $F(x) = (f_1(x), f_2(x), \cdots, f_m(x))^{\mathrm{T}}$，其中 $f_i(x)$，$i = 1,2,\cdots,m$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数. 作为光滑函数雅可比的推广，**克拉克广义雅可比**定义如下：

$$\partial F(x) = \mathrm{co}\{\lim_{x_i\to x} JF(x_i) \mid F(x) \text{ 在 } x = x_i \text{ 处可微}, x_i \to x\}. \tag{2-14}$$

易见,克拉克广义雅可比是 $m \times n$ 阶矩阵所构成的集合. 当 $\boldsymbol{F}(\boldsymbol{x})$ 是连续可微时, $\partial \boldsymbol{F}(\boldsymbol{x}) = \{\boldsymbol{JF}(\boldsymbol{x})\}$.

定义范数 $\| z \|_{m\times n} = \left(\sum_{i=1}^{m} \| z_i \|^2 \right)^{1/2}$,其中 z_i 为 z 的第 i 个列向量.

命题 13 设 $f_i(\boldsymbol{x}), i = 1,2,\cdots,m$ 是 $\mathbf{R}^n$ 上局部利普希茨函数,记 $\boldsymbol{F}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}))^{\mathrm{T}}$,则有下述结论:

1° $\partial \boldsymbol{F}(\boldsymbol{x})$ 是 $\mathbf{R}^{m\times n}$ 上的非空凸紧集;

2° 集值映射 $\boldsymbol{x} \to \partial \boldsymbol{F}(\boldsymbol{x})$ 是上半连续的.

定理 9 设 $f(\boldsymbol{x}) = g(\boldsymbol{F}(\boldsymbol{x}))$,其中 $\boldsymbol{F}: \mathbf{R}^n \to \mathbf{R}^m$ 和 $g: \mathbf{R}^m \to \mathbf{R}^1$ 是局部利普希茨的,则有下述链锁法则

$$\partial f(\boldsymbol{x}) \subset \mathrm{co}\{\partial \boldsymbol{F}(\boldsymbol{x})^{\mathrm{T}} \partial g(\boldsymbol{F}(\boldsymbol{x}))\}, \tag{2-15}$$

特别当 g 是连续可微函数时,有

$$\partial f(\boldsymbol{x}) = \partial \boldsymbol{F}(\boldsymbol{x})^{\mathrm{T}} \nabla g(\boldsymbol{F}(\boldsymbol{x})). \tag{2-16}$$

近年来,作为光滑函数雅可比的一种推广,主要是为求解非光滑方程组的需要,有人定义了集合

$$\partial_* \boldsymbol{F}(\boldsymbol{x}) = \partial f_1(\boldsymbol{x}) \times \partial f_2(\boldsymbol{x}) \times \cdots \times \partial f_m(\boldsymbol{x}). \tag{2-17}$$

从计算角度看,$\partial_* \boldsymbol{F}(\boldsymbol{x})$ 容易实现,但在链锁法则(例如公式(2-15))和隐函数定理中使用 $\partial \boldsymbol{F}(\boldsymbol{x})$,就要较 $\partial_* \boldsymbol{F}(\boldsymbol{x})$ 合理,易见,$\partial \boldsymbol{F}(\boldsymbol{x}) \subset \partial_* \boldsymbol{F}(\boldsymbol{x})$.

定理 10 设 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$, $\boldsymbol{F}$ 是在包含线段 $[\boldsymbol{x}, \boldsymbol{y}]$ 的某一开集上的向量局部利普希茨函数,则存在线段 $(\boldsymbol{x}, \boldsymbol{y})$ 中一点 $\boldsymbol{\xi}$,使得

$$\boldsymbol{F}(\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x}) \in \langle \partial \boldsymbol{F}(\boldsymbol{\xi}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

2.4 广义梯度的几何理论

设 C 为 $\mathbf{R}^n$ 中非空闭集,$d_C(x)$ 为点到集合 C 的距离函数,集合

$$T_C(\boldsymbol{x}) = \{\boldsymbol{h} \in \mathbf{R}^n \mid d^\circ_C(\boldsymbol{x}; \boldsymbol{h}) = 0\} \tag{2-18}$$

称为 C 在 $\boldsymbol{x}$ 处的**克拉克切锥**.

由于 $d_C(\boldsymbol{x}, \cdot)$ 是次线性函数,不难验证,$T_C(\boldsymbol{x})$ 是 $\mathbf{R}^n$ 中闭凸锥,当然也是闭凸集. 于是满足:对任意 $\boldsymbol{h}_1, \boldsymbol{h}_2 \in T_C(\boldsymbol{x})$,有 $\boldsymbol{h}_1 + \boldsymbol{h}_2 \in T_C(\boldsymbol{x})$. 当 C 为光滑超曲面时,$T_C(\boldsymbol{x})$ 为 C 在 $\boldsymbol{x}$ 处的切超平面.

$T_C(\boldsymbol{x})$ 的负对偶锥

$$N_C(\boldsymbol{x}) = \{\boldsymbol{v} \in \mathbf{R}^n \mid \langle \boldsymbol{v}, \boldsymbol{h} \rangle \leqslant 0, \forall \boldsymbol{h} \in T_C(\boldsymbol{x})\} \tag{2-19}$$

称为 C 在 $\boldsymbol{x}$ 处的**克拉克法锥**.

克拉克法锥 $N_C(\boldsymbol{x})$ 也是闭凸锥,当 C 为光滑超曲面时,$N_C(\boldsymbol{x})$ 就是 C 在 $\boldsymbol{x}$ 处的法向量.

切锥和法锥可分别表示为下述形式:

$$T_C(\boldsymbol{x}) = \{\boldsymbol{h} \in \mathbf{R}^n \mid t_i \to 0^+, \boldsymbol{x}_i \to \boldsymbol{x}, \boldsymbol{h}_i \to \boldsymbol{h}, \boldsymbol{x}_i + t_i \boldsymbol{h}_i \in C\}, \tag{2-20}$$

$$N_C(x) = \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)}. \tag{2-21}$$

证明(2-21)式：设 $z \in \partial d_C(x)$，根据克拉克广义梯度定义，有 $\langle z, h \rangle \leq d^\circ_C(x; h)$，$\forall h \in \mathbf{R}^n$. 对任意 $h \in T_C(x)$，由克拉克切锥定义，有 $d^\circ_C(x; h) = 0$，于是 $\langle z, h \rangle \leq 0$，故 $z \in N_C(x)$. 由于 $N_C(x)$ 是闭凸锥，有

$$\overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)} \subset N_C(x). \tag{2-22}$$

另一方面，设 $z \in N_C(x)$，由克拉克切锥和法锥定义，有 $\langle z, h \rangle \leq 0 = d^\circ_C(x; h)$，$\forall h \in T_C(x)$. 设 $h \bar{\in} T_C(x)$，此时 $h \neq 0$，于是有

$$d^\circ_C(x; h) = \overline{\lim_{\substack{x' \to x \\ t \to 0^+}}} \frac{d_C(x' + th) - d_C(x')}{t}$$

$$\geq \overline{\lim_{t \to 0^+}} \frac{1}{t} d_C(x + th) \geq 0.$$

当 $z = 0$ 时，则 $\langle z, h \rangle = 0 \leq d^\circ_C(x; h)$；当 $z \neq 0$ 时，选取 $\lambda_{z,h} = d^\circ_C(x; y) / \|z\| \|y\| \geq 0$，此时 $\lambda_{z,h} \langle z, h \rangle \leq \lambda_{z,h} \|z\| \|h\| = d^\circ_C(x; y)$. 由此可知，对任意 $z \in N_C(x)$，$y \in \mathbf{R}^n$，有 $\lambda_{z,h} \geq 0$，使得 $\lambda_{z,h} \langle z, y \rangle \leq d^\circ_C(x; h)$（当 $y \in T_C(x)$ 或 $z = 0$ 时，取 $\lambda_{z,h} = 1$），于是有 $z \in \bigcup_{\lambda \geq 0} \lambda \partial d_C(x)$，即

$$N_C(x) \subset \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)}. \tag{2-23}$$

综合(2-22)式和(2-23)式，得(2-21)式. 得证.

克拉克切锥和法锥还有下述性质：

(1) 当 C 为凸闭集时，$T_C(x) = \{h \in \mathbf{R}^n \mid d'_C(x; h) = 0\}$；

(2) 当 $x \in \text{int} C$ 时，$T_C(x) = \mathbf{R}^n$，$N_C(x) = \{0\}$.

设 $f(x)$ 是 $\mathbf{R}^n$ 上的实函数，$\mathbf{R}^n \times \mathbf{R}^1$ 中子集

$$\text{epi} f = \{(x, r) \in \mathbf{R}^n \times \mathbf{R}^1 \mid f(x) \leq r\} \tag{2-24}$$

称为 $f(x)$ 的**上图**. 设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，那么函数 $f^\circ(x; \cdot)$ 的上图是包含原点的凸锥.

命题 14 设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，则有

$$\text{epi} f^\circ(x; \cdot) = T_{\text{epi}} f(x, f(x)),$$

$$\partial f(x) = \{\xi \in \mathbf{R}^n \mid (\xi, -1) \in N_{\text{pei}} f(x, f(x))\}.$$

设 $f(x)$ 是 $\mathbf{R}^n$ 上的实函数，则集合

$$\text{lev} f(x) = \{y \in \mathbf{R}^n \mid (y, f(x)) \in \text{epi} f(x)\}$$

称为 $f(x)$ 在 x 处的**水平集**.

命题 15 设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，如果在点 x 处有 $0 \bar{\in} \partial f(x)$，则有

$$\{U \in \mathbf{R}^n \mid (U, 0) \in \text{epi} f^\circ(x; \cdot)\} \subset T_{\text{lev} f(x)}(x). \tag{2-25}$$

进一步，如果 $f(x)$ 在点 x 处是正则的，那么在(2-25)式中等式成立.

命题 16 设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数，如果在点 x 处有 $0 \bar{\in} \partial f(x)$，则有

$$N_{\mathrm{lev}f(x)}(x) \subset \bigcup_{\lambda \geq 0} \lambda \partial f(x). \tag{2-26}$$

进一步，如果 $f(x)$ 在点 x 处是正则的，则(2-26) 式中等式成立.

3　非光滑最优性理论

3.1　变 分 原 理

定理 1　设 $f(x)$ 是集合 K 上的局部利普希茨函数，当 $x \in K$ 满足

$$f(x) = \min_{y \in K} f(y),$$

则存在常数 $\alpha > 0, \delta > 0$，使得 $c \geq \alpha$ 时有

$$f(x) + cd_K(x) \leq f(y) + cd_K(y), \quad \forall y \in B(x,\delta), \tag{3-1}$$

进一步有

$$0 \in \partial(f + cd_K)(x) \subset \partial f(x) + N_K(x). \tag{3-2}$$

证明　由 f 的局部利普希茨性质，存在 $\delta' > 0, \alpha > 0$，使得 $c \geq \alpha$ 时，有

$$| f(x_1) - f(x_2) | \leq c \| x_1 - x_2 \|, \forall x_1, x_2 \in B(x,\delta'). \tag{3-3}$$

取 $\varepsilon \in (0,1)$ 充分小及 $\delta \in (0,\delta'/3)$，则对于 $y \in B(x,\delta)$，存在 $y_\varepsilon \in K$，使得 $\| y - y_\varepsilon \| \leq (1+\varepsilon) d_K(y)$. 由于 $\| x - y \| \leq \delta < \delta'/3$，故

$$\begin{aligned} \| x - y_\varepsilon \| &\leq \| x - y \| + \| y - y_\varepsilon \| \\ &\leq \| x - y \| + (1+\varepsilon) d_K(y) \\ &\leq (2+\varepsilon) \| x - y \| < \delta'. \end{aligned}$$

因此，由(3-3) 式可得

$$f(y_\varepsilon) \leq f(y) + c \| y - y_\varepsilon \| \leq f(y) + c(1+\varepsilon) d_K(y).$$

另一方面，由 $y_\varepsilon \in K$ 和 $f(x) \leq f(y_\varepsilon)$ 以及 x 为极小点，有

$$f(x) = f(x) + cd_K(x) \leq f(y) + c(1+\varepsilon) d_K(y).$$

由 ε 的任意性，(3-1) 式成立，进一步(3-2) 式也成立.

上述定理的意义在于将约束优化问题转化为无约束优化问题，其中乘子只要充分大，而不必趋向于无穷大.

定理 2　设 g 是 $\mathbf{R}^n$ 上局部利普希茨函数，给定 $x \in \mathbf{R}^n$，假定 $0 \bar{\in} \partial g(x)$，定义集合

$$K = \{ y \in \mathbf{R}^n \mid g(y) \leq g(x) \},$$

则

$$\{ h \in \mathbf{R}^n \mid g^\circ(x;h) \leq 0 \} \subset T_K(x). \tag{3-4}$$

证明　注意到存在 $h_0 \in \mathbf{R}^n$，使得 $g^\circ(x;h) < 0$，否则，$g^\circ(x;h) \geq 0, \forall h \in \mathbf{R}^n$，根据凸紧集和支撑函数之间的关系，有 $0 \in \partial g(x)$，与假设条件矛盾. 设 $h \in \{ h \in \mathbf{R}^n \mid g^\circ(x;h) \leq 0 \}$，对任意 $\varepsilon > 0$，有

$$g^\circ(x;h+\epsilon h_0) \leqslant g^\circ(x;h) + \epsilon g^\circ(x;h_0) < 0.$$

由集合 $T_K(x)$ 的闭性及上式,只要能够证明

$$\{h \in \mathbf{R}^n \mid g^\circ(x;h) < 0\} \subset T_K(x),$$

即得(3-4)式.

现假定 h 满足 $g^\circ(x;h) = -\delta$,其中 $\delta > 0$,则当 y 充分接近 x 点和 t 充分小时,有 $g(y+th) - g(y) \leqslant -\delta t$,特别当 $\{x_k\} \subset K, x_k \to x, t_k \to 0^+$, k 充分大时,有

$$g(x_k + t_k h) \leqslant g(x_k) - \delta t_k \leqslant g(x) - \delta t_k,$$

从而 $x_k + t_k h \in K$. 再由(2-20)式可得 $h \in T_K(x)$,(3-4)式成立,定理得证.

推论 1　在定理2的假设条件下,有

$$N_K(x) \subset \bigcup_{\lambda \geqslant 0} \lambda \partial g(x).$$

推论 2　设 g 满足定理2中的假设条件,x 为问题

$$\min f(y),$$
$$\text{s.t.}\ g(y) \leqslant 0, \quad y \in \mathbf{R}^n$$

的最优解,则存在 $\lambda \geqslant 0$,使得

$$\mathbf{0} \in \partial f(x) + \lambda \partial g(x).$$

为研究一般形式的最优性理论(主要是含有等式约束条件的问题),需要引入下述重要定理:**埃克朗(Ekeland)变分原理**,这一定理在研究各种非凸极值问题中都起到重要作用.

定理 3(埃克朗变分原理)　设 V 是完备距离空间,d 是 V 上的距离,$F: V \to \mathbf{R}^1 \cup \{-\infty\}$ 是下有界下半连续函数,且不恒为 $+\infty$. 又设对给定 $\epsilon > 0$,点 $u \in V$ 满足

$$F(u) \leqslant \inf_{v \in V} F(v) + \varepsilon, \tag{3-5}$$

那么存在 $v \in V$,使得

$$F(v) \leqslant F(u), \tag{3-6a}$$
$$d(u,v) \leqslant 1, \tag{3-6b}$$
$$F(w) > F(v) - \epsilon d(v,w), \ \forall\, w \in V, w \neq v. \tag{3-6c}$$

证明　按下述原则定义序列 $\{u_n\}$:

1° 令 $u_0 = u$.

2° 设 $u_n \in V$ 已知,如果

$$F(w) > F(u_n) - \epsilon d(u_n, w), \ \forall\, w \in V, w \neq u_n,$$

令 $u_{n+1} = u_n$;否则,令

$$S_n = \{w \in V \mid F(w) \leqslant F(u_n) - \epsilon d(u_n, w)\},$$

取 $u_{n+1} \in S_n$,使其满足

$$F(u_{n+1}) - \inf_{v \in S_n} F(v) \leqslant \frac{1}{2}\{F(u_n) - \inf_{v \in S_n} F(v)\}.$$

按上述原则定义的序列 $\{u_n\}$ 称为柯西列. 事实上,根据 $\{u_n\}$ 定义,有

$$F(u_n) - F(u_{n+1}) \geqslant \epsilon d(u_n, u_{n+1}), n = 1,2,\cdots,\text{从而}$$

$$F(u_n) - F(u_p) \geqslant \epsilon d(u_n, u_p), \quad \forall\, p \geqslant n, \tag{3-7}$$

因此$\{F(u_n)\}$是不增序列;又 F 下有界,故序列$\{F(u_n)\}$有极限.这样,由(3-7)式可知,$d(u_n,u_p)$可任意小,也就是说$\{u_n\}$是柯西列.

V是完备空间,$\{u_n\}$的极限 $v\in V$.现在证明上述 v 满足(3-6a),(3-6b)和(3-6c)式要求.由 F 下半连续和$\{F(u_n)\}$不增,可知

$$F(v)\leqslant\lim_{n\to\infty}F(u_n)\leqslant F(u_0)=F(u),$$

即(3-6a)式成立.同时,由

$$\varepsilon d(u,u_p)\leqslant F(u)-F(u_p)\leqslant F(u)-\inf_{v\in V}F(v)\leqslant\varepsilon,$$

可得

$$d(u,v)=\lim_{p\to\infty}d(u,u_p)\leqslant 1,$$

即(3-6b)式成立.另外,如果存在 $w\neq v$,使得 $F(w)\leqslant F(v)-\varepsilon d(v,w)$,由 $v=\lim\limits_{p\to\infty}u_p$ 和(3-7)式得

$$F(w)+\varepsilon d(v,w)\leqslant F(v)\leqslant F(u_n)-\varepsilon d(u_n,v).$$

进一步,有

$$\begin{aligned}F(w)&\leqslant F(u_n)-\varepsilon d(u_n,v)-\varepsilon d(v,w)\\&\leqslant F(u_n)-\varepsilon d(u_n,w).\end{aligned}$$

因此,对任意 $w\in S_n$,在下式

$$F(w)\geqslant\inf_{v\in S_n}F(v)\geqslant 2F(u_{n+1})-F(u_n)$$

中,关于 n 取极限,得右端极限大于等于 $F(v)$,这就与 w 的定义矛盾,故(3-6c)式成立.

推论3 设 F 是闭集$K\subset\mathbf{R}^n$上的下有界局部利普希茨函数,那么对于给定的 $\varepsilon>0$ 和满足

$$F(u)\leqslant\inf_{x\in\mathbf{R}^n}F(x')+\varepsilon$$

的点 $u\in K$,存在 $v\in K$,使得 $F(v)\leqslant F(u)$,$\|u-v\|\leqslant\sqrt{\varepsilon}$,$0\in\partial(F+cd_K)(v)+\sqrt{\varepsilon}B(0,1)$.

3.2 拉格朗日乘子法

本节将利用拉格朗日乘子给出约束利普希茨规划的最优性条件.下面将分具有不等式约束和具有等式与不等式约束两种情况来讨论.

首先考虑下述不等式约束问题

$$(\mathrm{P}_1)\qquad \begin{aligned}&\min f(x),\\&\text{s.t.}\quad g_i(x)\leqslant 0,\ i=1,2,\cdots,p,\ x\in K,\end{aligned}$$

其中 K 是 $\mathbf{R}^n$ 中某一集合,$f(x)$,$g_i(x)$是$\mathbf{R}^n$上的局部利普希茨函数.

定理4 设 $x_0\in K$ 为问题(P_1)的解,那么存在不全为零的常数 $\lambda_i\geqslant 0$,$i=0,1,2,\cdots,p$,使得

$$0 \in \lambda_0 \partial f(x_0) + \sum_{i=1}^{p} \lambda_i \partial g_i(x_0) + N_K(x_0), \tag{3-8a}$$

$$\lambda_i g_i(x_0) = 0, \quad i = 1,2,\cdots,p. \tag{3-8b}$$

进一步,如果存在 $h_0 \in \mathbf{R}^n$,使得

$$g^\circ_i(x_0;h_0) < 0, \quad i = 1,2,\cdots,p, \tag{3-9a}$$

则 $\lambda_0 = 1$.

证明 不难验证问题(P_1)等价于下述问题

(P_2) $\min\limits_{x\in K} F(x) = \max\{f(x) - f(x_0), g_i(x), i = 1,2,\cdots,p\}$.

由本章3.1节定理1,有

$$0 \in \partial F(x_0) + N_K(x_0).$$

再根据极大值函数克拉克广义梯度公式(2.2节命题10),可立刻得到(3-8a)式和(3-8b)式. 当(3-9a)式成立时,令 $G(x) = \max\limits_{1\leqslant i\leqslant p} g_i(x)$, $K_1 = \{x \mid \in \mathbf{R}^n \mid G(x) \leqslant G(x_0)\}$. 不妨假定 $G(x_0) = 0$(否则约束条件在点 x_0 处附近不起作用),这时(P_1)等价于下述问题

(P_3) $$\min_{x\in K_1\cap K} f(x),$$

于是,再由本章3.1节定理1,得

$$0 \in \partial f(x_0) + N_{K_1\cap K_2}(x_0).$$

不难验证

$$N_{K\cap K_1}(x_0) \subset N_K(x_0) + N_{K_1}(x_0).$$

根据(3-9)式,可得 $G^\circ(x_0;h_0) < 0$,从而 $0 \bar{\in} \partial G(x_0)$.

因此,由本章3.1节推论1可得

$$N_{K_1}(x_0) \subset \bigcup_{\lambda\geqslant 0} \lambda \partial G(x_0) = \bigcup_{\lambda\geqslant 0} \lambda \mathrm{co}\{\partial g_i(x_0) \mid i \in I(x_0)\},$$

其中

$$I(x^0) = \{i \in (1,2,\cdots,p) \mid g_i(x_0) = 0\}.$$

从而,在(3-8a)式和(3-8b)式中 $\lambda_0 = 1$.

(3-8a)式和(3-8b)式称为问题(P_1)的**弗利茨·约翰(Fritz John)必要性条件**,当式中 $\lambda_0 = 1$ 时称为**库恩-塔克(Kuhn-Tucker)必要性条件**. 满足(3-8a)式和(3-8b)式的点称为问题(P_1)的**弗利茨·约翰稳定点**;当 $\lambda_0 = 1$ 时满足(3-8a)式和(3-8b)式的点称为**库恩-塔克稳定点**.

定理4及证明方法不能直接推广到具有等式的约束问题. 为研究具有等式约束最优性条件,需要使用埃克朗变分原理. 考虑下述问题

(P_4)
$$\begin{aligned} &\min f(x), \\ &\text{s.t. } g_i(x) \leqslant 0, \quad i = 1,2,\cdots,p, \\ &\quad h_j(x) = 0, \quad j = 1,2,\cdots,q, \\ &\quad x \in K, \end{aligned}$$

其中 K 为 $\mathbf{R}^n$ 中的某一集合,$f(x),g_i(x),h_j(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数.

定理 5 设 x_0 是问题(P_4)的最优解,则对于充分大的 $c>0$,存在不全为零的常数 $\lambda_0,\lambda_1,\cdots,\lambda_p\geqslant 0,\mu_1,\cdots,\mu_q$,使得

$$0\in\lambda_0\partial f(x_0)+\sum_{i=1}^{p}\lambda_i\partial g_i(x_0)+\sum_{j=1}^{q}\mu_j\partial h_j(x_0)+cN_K(x_0),\tag{3-9b}$$

$$\lambda_i g_i(x_0)=0,\quad i=1,2,\cdots,p.\tag{3-9c}$$

证明 令 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_p)\in\mathbf{R}_+^p,\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_q)\in\mathbf{R}^q,\boldsymbol{g}(x)=(g_1(x),g_2(x),\cdots,g_p(x)),\boldsymbol{h}(x)=(h_1(x),h_2(x),\cdots,h_q(x))$.定义集合

$$T=\{t=(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathbf{R}^{1+p+q}\mid\lambda_0\geqslant 0,\boldsymbol{\lambda}\in\mathbf{R}_+^p,\|(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu})\|=1\}.$$

给定 $\varepsilon>0$,令

$$\begin{aligned}F_\varepsilon(x)&=\max\{\langle(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu}),(f(x)-f(x_0)+\varepsilon,\boldsymbol{g}(x),\boldsymbol{h}(x))\rangle\mid(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu})\in T\}\\&=\max\{\lambda_0(f(x)-f(x_0)+\varepsilon)+\sum_{i=1}^{p}\lambda_i g_i(x)+\\&\qquad\sum_{j=1}^{q}\mu_j h_j(x)\mid(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu})\in T\}.\end{aligned}$$

易见,$F_\varepsilon(x)$ 是 $\mathbf{R}^n$ 上局部利普希茨函数,且有 $F_\varepsilon(x_0)=\varepsilon$.另一方面,$F_\varepsilon(x)>0$,$\forall x\in K$,若不然,存在 $x_1\in K$,使得 $F_\varepsilon(x_1)\leqslant 0$,此时有 $g_i(x_1)\leqslant 0,h_j(x_1)=0$,$f(x_1)\leqslant f(x_0)-\varepsilon$,这与 x_0 是最优解矛盾.

综合以上分析可得,x_0 满足 $F_\varepsilon(x_0)\leqslant\inf_{x\in K}F_\varepsilon(x)+\varepsilon$.由埃克朗变分原理的推论可知,存在 $x_\varepsilon\in K$,使得

$$\|x_0-x_\varepsilon\|\leqslant\sqrt{\varepsilon},$$

$$0\in\partial(F_\varepsilon(x_\varepsilon)+cd_K(x_\varepsilon))+\sqrt{\varepsilon}B(0,1).$$

进一步,有

$$0\in\partial(F_\varepsilon(x_\varepsilon)+cd_K(x_\varepsilon))+\sqrt{\varepsilon}B(0,1).$$

由广义梯度运算可知,存在不全为零的常数(依赖于 ε)$\lambda_0(\varepsilon),\lambda_1(\varepsilon),\cdots,\lambda_p(\varepsilon)\geqslant 0,\mu_1(\varepsilon),\cdots,\mu_q(\varepsilon)$,使得

$$\begin{aligned}0\in\partial(\lambda_0(\varepsilon)f(x_\varepsilon)+\sum_{i=1}^{p}\lambda_i(\varepsilon)g_i(x_\varepsilon)+\sum_{j=1}^{q}\mu_j(\varepsilon)h_j(x_\varepsilon)+\\cd_K(x_\varepsilon))+\sqrt{\varepsilon}B(0,1).\end{aligned}\tag{3-10}$$

在(3-10)式中令 $\varepsilon\to 0$,则有 $x_\varepsilon\to x_0$.根据 T 是有界闭集可知,存在 $\varepsilon_n\to 0$ 使得 $\lambda_0(\varepsilon_n)\to\lambda_0,\lambda_i(\varepsilon_n)\to\lambda_i,\mu_j(\varepsilon_n)\to\mu_j$,且有 $\|(\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu})\|=1$.再根据广义梯度上半连续性质,可得

$$\begin{aligned}0&\in\partial(\lambda_0 f(x_0)+\sum_{i=1}^{p}\lambda_i g_i(x_0)+\sum_{j=1}^{q}\mu_j h_j(x_0)+cd_K(x_0))\\&\subset\lambda_0\partial f(x_0)+\sum_{i=1}^{p}\lambda_i\partial g_i(x_0)+\sum_{j=1}^{q}\mu_j\partial h_j(x_0)+N_K(x_0).\end{aligned}$$

定理得证.

定理5是具有等式与不等式约束条件的利普希茨规划的拉格朗日法则.(3-9b)和(3-9c)称为弗利茨·约翰条件;当 $\lambda_0=1$ 时称为库恩-塔克条件.

3.3　灵敏度和平稳性分析

本节讨论约束条件摄动时优化问题的灵敏度.稳定性是灵敏度分析中的一个重要概念.

命题1　设 S 是 $\mathbf{R}^n$ 上的有界凸集,T 是 $\mathbf{R}^m$ 上的凸集,$A:S\to Y$ 是线性映射,$f:S\to\mathbf{R}^1$ 是利普希茨函数.考虑下述摄动优化问题

$$(\mathrm{P}_y)\qquad \begin{aligned}&\min f(x),\\&Ax\in T+y,\\&x\in S'.\end{aligned}$$

定义**最优值函数**

$$V(y)=\inf\{f(x)\mid x\in S,Ax\in T+y\}.$$

如果存在 $\epsilon>0$,使得 $\epsilon B_y\subset AS-T$,那么,V 在 $\mathbf{0}$ 点附近是利普希茨函数.

命题2　在本章3.2节定理4假设条件下又假定 K 是凸集,$g_i(x)$ 是 K 上凸函数且满足斯莱特(L.J.Slater)条件:存在 $x_0\in K$,使得 $g_i(x_0)<0,i=1,2,\cdots,p$,那么,对于 $y=(y_1,y_2,\cdots,y_p)^{\mathrm{T}}\in\mathbf{R}^q$,最优值函数

$$V(y)=\inf\{f(x)\mid g_i(x)+y_i\leqslant 0,i=1,2,\cdots,p,x\in K\}$$

在 $\mathbf{0}$ 点附近是利普希茨函数.

对于一般具有等式和不等式约束条件的摄动问题

$$(\mathrm{P}_{ab})\qquad \begin{aligned}&\min f(x),\\&g_i(x)+a_i\leqslant 0,\quad i=1,2,\cdots,p,\\&h_j(x)+b_j=0,\quad j=1,2,\cdots,q,\\&x\in K.\end{aligned}$$

定义**最优值函数** $V:\mathbf{R}^p\times\mathbf{R}^q\to\mathbf{R}\cup\{\pm\infty\}$ 如下:

$$\begin{aligned}V(a,b)=\inf\{f(x)\mid &g_i(x)+a_i\leqslant 0,i=1,2,\cdots,p,\\&h_j(x)+b_j=0,j=1,2,\cdots,q,x\in K\}.\end{aligned}$$

如果 $V(0,0)$ 有限,且

$$\varlimsup_{\substack{a\to0\\b\to0}}\frac{V(a,b)-V(0,0)}{\|(a,b)\|}>-\infty,$$

那么称问题(P_{00})是平稳的.

命题3　假设问题(P_{00})是平稳的,x_0 是其最优解,则对某个 $M>0$,x_0 是函数

$$f(x)+M\max\{g_i(x),|h_j(x)|\mid i=1,2,\cdots,p,j=1,2,\cdots,q\}$$

在 K 上的局部极小点.

命题4　设问题(P_{00})是平稳的,那么在它的拉格朗日乘子法中 $\lambda_0=1$.

4　其他几种方向导数、微分及最优性理论

4.1　拟可微函数及其最优性理论

拟可微函数首先是由苏联学者普谢尼奇奈(Pschenichny)于20世纪60年代末提出来的.设$f(\boldsymbol{x})$是$\mathbf{R}^n$上的方向可微函数,对于$\boldsymbol{x}\in\mathbf{R}^n$,若存在凸紧集$\partial f(\boldsymbol{x})\subset\mathbf{R}^n$,使得$f(\boldsymbol{x})$在$\boldsymbol{x}$点的方向导数可表示为

$$f'(\boldsymbol{x};\boldsymbol{h})=\max_{\boldsymbol{v}\in\partial f(\boldsymbol{x})}\langle\boldsymbol{v},\boldsymbol{h}\rangle,\forall\boldsymbol{h}\in\mathbf{R}^n,\tag{4-1}$$

则称$f(\boldsymbol{x})$在$\boldsymbol{x}$点处是拟可微的,$\partial f(\boldsymbol{x})$称为**次微分**.不难看出,$f(\boldsymbol{x})$是拟可微的,当且仅当它的方向导数关于方向是凸函数.显然,凸函数,光滑函数的极大值函数是拟可微的.但是,在此定义下的拟可微函数类还不够广泛,例如当$f(\boldsymbol{x})$是拟可微的,$-f(\boldsymbol{x})$一般说来不再是拟可微的(除非$f(\boldsymbol{x})$本身是可微的).换句话说,拟可微函数类不能构成线性空间.

为讨论更一般的不可微函数,季米雅诺夫(Demyanov).鲁宾诺夫(Rubinov)和波雅柯娃(Polyakova)将普谢尼奇奈的**拟可微函数**定义进行了推广,定义拟可微函数如下:

设$f(\boldsymbol{x})$是$\mathbf{R}^n$上的方向可微函数,若其方向导数可表示为

$$f'(\boldsymbol{x};\boldsymbol{h})=\max_{\boldsymbol{v}\in\underline{\partial}f(\boldsymbol{x})}\langle\boldsymbol{v},\boldsymbol{h}\rangle+\min_{\boldsymbol{w}\in\overline{\partial}f(\boldsymbol{x})}\langle\boldsymbol{w},\boldsymbol{h}\rangle,\forall\boldsymbol{h}\in\mathbf{R}^n,\tag{4-2}$$

其中$\underline{\partial}f(\boldsymbol{x}),\overline{\partial}f(\boldsymbol{x})$是$\mathbf{R}^n$中非空凸紧集,则称$f(\boldsymbol{x})$是拟可微的.

有序集合对$\mathrm{D}f(\boldsymbol{x})=[\underline{\partial}f(\boldsymbol{x}),\overline{\partial}f(\boldsymbol{x})]$称为**拟微分**.$\underline{\partial}f(\boldsymbol{x})$和$\overline{\partial}f(\boldsymbol{x})$分别称为**次微分和超微分**.特别地,当$\overline{\partial}f(\boldsymbol{x})=\{0\}$时,称$f(\boldsymbol{x})$是次可微的;当$\underline{\partial}f(\boldsymbol{x})=\{0\}$时,称$f(\boldsymbol{x})$是超可微的.

由拟可微函数的定义不难看出,拟微分不是唯一确定的.事实上,若$[\underline{\partial}f(\boldsymbol{x}),\overline{\partial}f(\boldsymbol{x})]$是$f(\boldsymbol{x})$的一个拟微分,则对任意凸紧集$A\subset\mathbf{R}^n$,有$[\underline{\partial}f(\boldsymbol{x})+A,\overline{\partial}f(\boldsymbol{x})-A]$也是$f(\boldsymbol{x})$的一个拟微分.通常用$\mathscr{D}f(\boldsymbol{x})$记$f(\boldsymbol{x})$的全体拟微分集合.

拟可微函数和局部利普希茨函数,是两类互不包含的不可微函数类.许多常见的不可微函数都是拟可微的.

例1　设$f(\boldsymbol{x})=f_1(\boldsymbol{x})-f_2(\boldsymbol{x})$,其中,$f_1(\boldsymbol{x}),f_2(\boldsymbol{x})$均是$\mathbf{R}^n$上的凸函数,$f(\boldsymbol{x})$是$\mathbf{R}^n$上的拟可微函数,$[\partial f_1(\boldsymbol{x}),-\partial f_2(\boldsymbol{x})]$是一个拟微分.

例2　设$f(\boldsymbol{x})=\max\limits_{i\in I}f_i(\boldsymbol{x})+\min\limits_{j\in J}g_j(\boldsymbol{x})$,其中$f_i(\boldsymbol{x})$和$g_j(\boldsymbol{x})$均是$\mathbf{R}^n$上连续可微函数,$I,J$为有限指标集,定义指标集

$$I(\boldsymbol{x})=\{i\in I\mid f_i(\boldsymbol{x})=\max_{i\in I}f_i(\boldsymbol{x})\},$$

$$J(\boldsymbol{x})=\{j\in J\mid g_j(\boldsymbol{x})=\min_{j\in J}g_j(\boldsymbol{x})\}.$$

记$\underline{\partial}f(\boldsymbol{x})=\mathrm{co}\{\nabla f_i(\boldsymbol{x})\mid i\in I(\boldsymbol{x})\},\overline{\partial}f(\boldsymbol{x})=\mathrm{co}\{\nabla g_j(\boldsymbol{x})\mid j\in J(\boldsymbol{x})\}$,不难验证,

$[\underline{\partial}f(x),\bar{\partial}f(x)]$ 是 $f(x)$ 的一个拟微分.

拟可微函数及其拟微分有下述性质:

1° 设 $f(x)$ 是拟可微的,则对任意常数 c, $f_1(x)=cf(x)$ 是拟可微的,且有

$$\underline{\partial}f_1(x)=\begin{cases}c\,\underline{\partial}f(x), & \text{当 } c\geqslant 0,\\ c\bar{\partial}f(x), & \text{当 } c<0;\end{cases}$$

$$\bar{\partial}f_1(x)=\begin{cases}c\bar{\partial}f(x), & \text{当 } c\geqslant 0,\\ c\,\underline{\partial}f(x), & \text{当 } c<0.\end{cases}$$

2° 设 $f_1(x)$ 和 $f_2(x)$ 是拟可微的,则函数 $f(x)=f_1(x)+f_2(x)$ 是拟可微的,且有

$$\underline{\partial}f(x)=\underline{\partial}f_1(x)+\underline{\partial}f_2(x),$$
$$\bar{\partial}f(x)=\bar{\partial}f_1(x)+\bar{\partial}f_2(x).$$

3° 设 $f(x)=\max\limits_{i\in I}f_i(x)$,其中 $f_i(x)$ 是拟可微的, I 为有限指标集,则 $f(x)$ 是拟可微的,且有

$$\underline{\partial}f(x)=\operatorname{co}\{\underline{\partial}f_i(x)-\sum_{j\in I(x)\setminus\{i\}}\bar{\partial}f_j(x)\mid i\in I(x)\},$$
$$\bar{\partial}f(x)=\sum_{i\in I(x)}\bar{\partial}f_i(x),$$

其中 $I(x)=\{i\in I\mid f_i(x)=\max\limits_{i\in I}f_i(x)\}$.

4° 设 $f(x)=\min\limits_{i\in I}f_i(x)$,其中 $f_i(x)$ 是拟可微的, I 为有限指标集,则 $f(x)$ 是拟可微的,且有

$$\underline{\partial}f(x)=\sum_{i\in I(x)}\underline{\partial}f_i(x),$$
$$\bar{\partial}f(x)=\operatorname{co}\{\bar{\partial}f_i(x)-\sum_{j\in I(x)\setminus\{i\}}\underline{\partial}f_j(x)\mid i\in I(x)\},$$

其中 $I(x)=\{i\in I\mid f_i(x)=\min\limits_{i\in I}f_i(x)\}$.

5° 设 $f_i(x)$, $i=1,2,\cdots,m$ 是 $\mathbf{R}^n$ 上的拟可微函数, F 是 $\mathbf{R}^m$ 上的连续可微函数,则

$$f(x)=F(f_1(x),f_2(x),\cdots,f_m(x))$$

是拟可微的.

下述命题和定理是有关拟可微优化的一阶最优性必要条件.

命题 1 设 $f(x)$ 是 $\mathbf{R}^n$ 上的拟可微函数,若 $x^*\in\mathbf{R}^n$ 是 $f(x)$ 的局部极小点,则有

$$-\bar{\partial}f(x^*)\subset\underline{\partial}f(x^*);$$

若 $x^{**}\in\mathbf{R}^n$ 是 $f(x)$ 的局部极大点,则有

$$-\underline{\partial}f(x^{**})\subset\bar{\partial}f(x^{**}).$$

下面考虑不等式约束问题

$$\text{(P)}\qquad \min f_0(x),$$
$$\text{s.t. } f_i(x)\leqslant 0,\quad i=1,2,\cdots,m,$$

其中 $f_i(x), i = 0,1,2,\cdots,m$ 是 $\mathbf{R}^n$ 上的拟可微函数.关于问题(P)的最优性条件有两种类型,其中一类是几何型条件,例如下述定理1;另一类是拉格朗日乘子型条件,例如下述定理2.

定理1　设 $x^* \in \mathbf{R}^n$ 是问题(P)的最优解,则有

$$-\sum_{i\in I(x^*)\cup\{0\}} \overline{\partial} f_i(x^*) \subset \mathrm{co}\{\underline{\partial} f_i(x^*) - \sum_{j\in I(x^*)\cup\{0\}\setminus\{i\}} \overline{\partial} f_j(x^*) \mid i \in I(x^*) \cup \{0\}\},$$

其中 $I(x^*) = \{i \mid 1 \leqslant i \leqslant m, f_i(x^*) = 0\}$.

定理2　设 x^* 是问题(P)的最优解,则对任意一组超微分 $w_i \in \overline{\partial} f_i(x^*), i = 0,1,2,\cdots,m$,存在一组不全为零的常数 $\lambda_i(w) \geqslant 0, i = 0,1,2,\cdots,m$(其中 $\lambda_i(w)$ 依赖于 $w, w = (w_0, w_1, w_2, \cdots, w_m)$),使得

$$\mathbf{0} \in \sum_{i=0}^{m} \lambda_i(w)(\underline{\partial} f_i(x^*) + w_i),$$

$$\lambda_i(w) f_i(x^*) = 0,\quad i = 1,2,\cdots,m.$$

利用拟微分可建立下述**中值定理**.

定理3(中值定理)　设 $x, y \in \mathbf{R}^n$, f 是在包含线段 $[x,y]$ 的某一开集上的拟可微函数,则存在线段 (x,y) 中一点 ξ,使得

$$f(y) - f(x) \in \langle \underline{\partial} f(\xi) + \overline{\partial} f(\xi), y - x\rangle.$$

4.2　迪尼导数和贝诺特微分

4.2.1　迪尼导数

迪尼上、下导数分别定义如下:

$$f_{\mathrm{D}}^{\uparrow}(x;h) = \overline{\lim_{t\to 0^+}} \frac{(f(x+th) - f(x))}{t},$$

$$f_{\mathrm{D}}^{\downarrow}(x;h) = \underline{\lim_{t\to 0^+}} \frac{(f(x+th) - f(x))}{t}.$$

容易看出,当 $f(x)$ 是局部利普希茨函数时,迪尼上、下导数都取有限值.

命题2　设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数,$x^* \in \mathbf{R}^n$ 是 $f(x)$ 的局部极小点,则有 $f_{\mathrm{D}}^{\downarrow}(x^*;h) \geqslant 0, \forall h \in \mathbf{R}^n$;若对任意 $h \in \mathbf{R}^n$,有 $f_{\mathrm{D}}^{\downarrow}(x^{**},h) > 0$,则 x^{**} 是 $f(x)$ 的严格局部极小点.

命题3　设 $f(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨函数,若 $f_{\mathrm{D}}^{\uparrow}(x;h)$ 对于固定 h 关于 x 点是上半连续的,则有 $f_{\mathrm{D}}^{\uparrow}(x;h) = f^{\circ}(x;h)$.

命题4　设 $f(x)$ 是包含线段 $[x,y]$ 的某一开集上的连续函数,记

$$m = \inf_{0\leqslant t\leqslant 1} f_{\mathrm{D}}^{\uparrow}(x + t(y-x)),$$

$$M = \sup_{0\leqslant t\leqslant 1} f_{\mathrm{D}}^{\uparrow}(x + t(y-x)),$$

则有

$$m\|y-x\| \leqslant f(y) - f(x) \leqslant M\|y-x\|.$$

4.2.2　贝诺特微分

设 $f(x)$ 是 $\mathbf{R}^n$ 上的方向可微函数，贝诺特次微分和超微分分别定义如下：

$$\partial_{\leqslant} f(x) = \{v \in \mathbf{R}^n \mid \langle v, h \rangle \leqslant f'(x;h), \forall h \in \mathbf{R}^n\},$$

$$\partial_{\geqslant} f(x) = \{w \in \mathbf{R}^n \mid \langle w, h \rangle \geqslant f'(x;h), \forall h \in \mathbf{R}^n\}.$$

贝诺特次微分和超微分是闭凸集，但不能保证非空．容易证明 $\partial_{\geqslant} f(x) = -\partial_{\leqslant}(-f)(x)$．

命题 5　设 $f(x)$ 是 $\mathbf{R}^n$ 上的方向可微函数，若 x^* 是 $f(x)$ 的局部极小点，则有 $0 \in \partial_{\leqslant} f(x^*)$；若 x^{**} 是 $f(x)$ 的局部极大点，则有 $0 \in \partial_{\geqslant} f(x^{**})$．

设 A, B 均为 $\mathbf{R}^n$ 中集合，定义集合运算

$$A \div B = \{h \in \mathbf{R}^n \mid \{h\} + V \subset U\}.$$

显然，$A \div B$ 仍为 $\mathbf{R}^n$ 中集合，但可能是空集．

命题 6　设 $f(x)$ 是 $\mathbf{R}^n$ 上的拟可微函数，$[\underline{\partial} f(x), \bar{\partial} f(x)]$ 是 $f(x)$ 的一个拟微分，则有

$$\underline{\partial} f(x) \div (-\bar{\partial} f(x)) = \partial_{\leqslant} f(x),$$

$$(-\bar{\partial} f(x)) \div (\underline{\partial} f(x)) = \partial_{\geqslant} f(x).$$

5　不可微优化算法

5.1　算法概况

本章所涉及的**不可微优化算法**，是基于次梯度(关于凸规划)或克拉克广义梯度(关于利普希茨规划)理论所构造、设计的不可微优化算法，而不包括非线性规划中的直接法(例如，单纯形法，坐标轮换法，随机搜索法等)．到目前为止，不可微优化算法大多数还仅仅是“概念性算法”，而非“可执行算法”．因为在不可微优化算法中，要求在每一个迭代点处都能够计算函数克拉克广义梯度中一个元素，对大多数不可微函数来讲，这一点是很难做到的，除非所考虑的不可微函数具有很特殊结构，例如，光滑函数的极大值函数等．

现考虑无约束优化问题

$$(\mathrm{P}_1) \qquad \min_{x \in \mathbf{R}^n} f(x),$$

其中 $f(x)$ 是 $\mathbf{R}^n$ 上的实函数．经典的求解无约束问题(P_1)算法的迭代公式如下：

$$x_{k+1} = x_k + \lambda_k h_k,$$

其中 $h_k \in \mathbf{R}^n$ 是**搜索方向**，一般要求它是 $f(x)$ 在 $x = x_k$ 处的**下降方向**；λ_k 是**步长**，满足

$$f(x_{k+1}) = f(x_k + \lambda_k h_k) = \min_{\lambda \geqslant 0} f(x_k + \lambda h_k).$$

点列$\{x_k\}_1^\infty$期望收敛于问题(P_1)的最优解.

在上述算法框架中,一个关键的问题是如何选取搜索方向h_k.当$f(x)$是连续可微函数时,可选取$h_k = -\nabla f(x_k)$,这就是通常所说的最速下降法.当$f(x)$是一般局部利普希茨函数时,下述定理说明利用克拉克广义梯度可确定它的一个下降方向.

定理 1 设$f(x)$是$\mathbf{R}^n$上的局部利普希茨函数,且$\mathbf{0} \bar{\in} \partial f(x)$,则存在$\xi^* \in \partial f(x)$满足

$$\|\xi^*\| = \min_{\xi\in\partial f(x)} \|\xi\|, \tag{5-1}$$

使得$h = -\xi$是$f(x)$在x点的一个下降方向,即存在$T > 0$,使得

$$f(x + th) < f(x), \forall t \in (0, T). \tag{5-2}$$

证明 $\partial f(x)$是有界闭凸集,故满足(5-1)式的ξ^*是存在的,从而,

$$\|\xi\|^2 \geqslant \|\xi^*\|^2, \quad \forall \xi \in \partial f(x).$$

又$\partial f(x)$是凸集,故对任意$\lambda \in (0,1)$,有

$$\xi^* + \lambda(\xi - \xi^*) \in \partial f(x),$$

从而

$$\|\xi^* + \lambda(\xi - \xi^*)\| = \|\xi^*\|^2 + 2\lambda\langle\xi^*, \xi - \xi^*\rangle + \lambda^2\|\xi - \xi^*\|^2 \geqslant \|\xi^*\|^2.$$

于是

$$\langle\xi^*, \xi - \xi^*\rangle \geqslant -\frac{\lambda}{2}\|\xi - \xi^*\|^2, \quad \forall \xi \in \partial f(x).$$

由$\lambda \in (0,1)$的任意性及ξ^*满足(5-1)式,则有

$$\langle\xi, h\rangle \leqslant -\|h\|^2, \quad \forall \xi \in \partial f(x). \tag{5-3}$$

另一方面,根据中值定理(2.2 节定理 7),有

$$f(x + th) - f(x) \in \langle\partial f(x'), th\rangle, \tag{5-4}$$

其中$x' \in (x, x + th)$.由广义梯度的上半连续性质,对充分小的t,有

$$\partial f(x') \subset \partial f(x) + \frac{\|h\|}{2}B(\mathbf{0}, 1). \tag{5-5}$$

由(5-3)式和(5-5)式,得

$$\forall \xi_1 \in \partial f(x'), \langle\xi_1, h\rangle \leqslant -\|h\| + \frac{1}{2}\|h\|^2 = -\frac{1}{2}\|h\|.$$

再根据(5-4)式,对充分小的t,有

$$f(x + th) - f(x) \leqslant -\frac{1}{2}t\|h\|.$$

定理得证.

利用定理 1 可以给出求解问题(P_1)的"广义最速下降法",其中$f(x)$只要求是局部利普希茨函数.

算法 1

(1) 给定初始点$x_0 \in \mathbf{R}^n$,令$k = 0$.

(2) 求$\xi_k^* \in \partial f(x_k)$,满足$\|\xi_k^*\| = \min_{\xi\in\partial f(x_k)} \|\xi\|$,

若 $\xi_k^* = 0$,则停止,否则转(3).

(3) 令 $h_k = -\xi_k$,求步长 λ_k 满足

$$f(x_k + \lambda_k h_k) = \min_{\lambda \geq 0} f(x_k + \lambda h_k).$$

(4) 令 $x_{k+1} = x_k + \lambda_k h_k$, $k = k+1$,转(2).

由定理1可知,上述算法确定是一个下降算法,但存在两个问题:其一是无法具体实现,由算法中(2)可知,要想实现算法,需要在每一迭代点处知道广义梯度集合 $\partial f(x_k)$;其二是收敛性无法保证.

5.2　不可微凸规划算法

5.2.1　次梯度法

在问题(P_1)中假定 $f(x)$ 是 $\mathbf{R}^n$ 中凸函数,则下述算法称为**次梯度法**.

算法2

(1) 选取数列 $\{\lambda_k\}$ 满足 $\lambda_k > 0$, $\lambda_k \to 0 (k \to \infty)$, $\sum_{k=0}^{\infty} \lambda_k = +\infty$.

(2) 给定初始点 $x_0 \in \mathbf{R}^n$,令 $k = 0$.

(3) 选取次微分集合 $\partial f(x_k)$ 中任意一个元素,记为 ξ_k,若 $\xi_k = \mathbf{0}$,则停止,否则记 $h_k = -\xi_k$,转(4).

(4) 令 $x_{k+1} = x_k + \lambda_k h_k / \| h_k \|$, $k = k+1$,转(3).

命题1　设问题(P_1)中 $f(x)$ 是凸函数,解集 $S^* = \{x \mid f(x) = f^* = \min_{x \in \mathbf{R}^n} f(x)\}$ 非空,如果 $x_k \bar{\in} S^*$,则对任意 $x^* \in S^*$, $\xi_k \in \partial f(x_k)$,必存在 $T_k > 0$,使得

$$\left\| x_k - \lambda \frac{1}{\| \xi_k \|} \xi_k - x^* \right\| < \| x_k - x^* \|, \forall \lambda \in (0, T_k). \tag{5-6}$$

证明　直接计算有

$$\left\| x_k - \lambda \frac{1}{\| \xi_k \|} \xi_k - x^* \right\|^2 = \| x_k - x^* \|^2 + 2\lambda \left(\frac{\xi_k}{\| \xi_k \|} \right)^{\mathrm{T}} (x^* - x_k) + \lambda^2. \tag{5-7}$$

由于 $\xi_k \in \partial f(x_k)$ 和 $x_k \bar{\in} S^*$,有

$$\xi_k^{\mathrm{T}} (x^* - x_k) \leq f(x^*) - f(x_k) < 0. \tag{5-8}$$

于是,可取

$$T_k = -2\xi_k^{\mathrm{T}} (x^* - x_k) / \| \xi_k \|,$$

由(5-8)式易见(5-6)式成立.

定理2　记 S^* 为问题(P_1)的解集,$d_{S^*}(x)$ 为 x 到 S^* 的距离,$\{x_k\}$ 是算法2产生的点列,则有

$$\lim_{k \to \infty} d_{S^*}(x_k) = 0.$$

证明　由 $f(x)$ 的凸性,必存在连续函数 $\delta(\varepsilon)$ 使得

$$f(x) \leqslant f^* + \epsilon$$

对任何 $d_{S^*}(x) \leqslant \delta(\epsilon), \delta(\epsilon) > 0 (\epsilon > 0)$ 成立. 对每个 k, 定义 $\epsilon_k = f(x_k) - f^* \geqslant 0$. 如果 $\epsilon_k > 0$, 则有

$$\begin{aligned}\| x_{k+1} - x^* \|^2 &= \| x_k - x^* \|^2 + \lambda_k^2 - 2\lambda_k (x_k - x^*)^T \xi_k / \| \xi_k \| \\ &= \| x_k - x^* \|^2 + \lambda_k^2 - 2\delta(\epsilon_k)\lambda_k - \\ &\quad 2\lambda_k \left(x_k - x^* - \delta(\epsilon_k) \frac{\xi_k}{\| \xi_k \|} \right)^T \xi_k / \| \xi_k \| \\ &\leqslant \| x_k - x^* \|^2 + \lambda_k^2 - 2\delta(\epsilon_k)\lambda_k. \end{aligned} \tag{5-9}$$

由上式可知

$$d_{S^*}^2(x_{k+1}) - d_{S^*}^2(x_k) \leqslant -\lambda_k(2\delta(\epsilon_k) - \lambda_k). \tag{5-10}$$

定义 $\delta(0) = 0$, 则上式对于一切 k 都成立. 对(5-10)式两边求和后易见

$$\lim_{k\to\infty} \delta(\epsilon_k) = 0,$$

于是有

$$\lim_{k\to\infty} d_{S^*}(x_{k+1}) = 0.$$

假设定理不真, 则必存在正常数 $\delta' > 0$ 和无穷多个 k 使得下式成立:

$$d_{S^*}(x_{k+1}) > d_{S^*}(x_k), \epsilon_k > \delta'. \tag{5-11}$$

由(5-10)式和(5-11)式可得 $2\delta(\epsilon_k) < \lambda_k$, 这与 $\epsilon_k > \delta'$ 矛盾. 定理得证.

给定 $\epsilon > 0$, 在算法 2 中选取 $\xi_k \in \partial_\epsilon f(x_k)$, 就得到 ϵ 次梯度法. 一般说来, 计算 ϵ 次微分 $\partial_\epsilon f(x_k)$ 中的一个元素要较计算次微分 $\partial f(x_k)$ 中的一个元素容易, 因此 ϵ 次梯度法更容易实现. 当然, 还可选取 $\epsilon_k \to 0, \epsilon_k > 0$, 在算法 2 中选取 $\xi_k \in \partial_{\epsilon_k} f(x_k)$, 得到的算法称为 ϵ_k 次梯度法. 可以证明 ϵ 次梯度法和 ϵ_k 次梯度法都是收敛的.

5.2.2 割平面法

对于 $\mathbf{R}^n$ 的凸函数 $f(x)$, 有

$$f(x) = \max_{y\in\mathbf{R}^n} \max_{\xi\in\partial f(y)} (f(y) + \xi^T(x - y)).$$

所以, 求解 $f(x)$ 极小就等价于求解下述问题

(P_2)

$$\begin{aligned}&\min\ v, \\ &\text{s.t.}\quad f(y) + \xi^T(x - y) \leqslant v, \forall y \in \mathbf{R}^n, \xi \in \partial f(y).\end{aligned}$$

假设 $x_i, i = 1,2,\cdots,k$ 是已有的迭代点, 在第 k 次迭代中求解问题

(P_3)

$$\begin{aligned}&\min\ v, \\ &\text{s.t.}\quad f(x_i) + \xi_i^T(x - x_i) \leqslant v, \quad i = 1,2,\cdots,k.\end{aligned}$$

显然, 线性规划问题(P_3)是问题(P_2)的一个逼近. 下面给出求解问题(P_1)的一个割平面算法.

算法 3

(1) 给定初始点 $x_1 \in S$, S 是一给定的凸多面体,令 $k = 1$.

(2) 计算次微分集合 $\partial f(x_k)$ 中任一元素,并记为 ξ_k,若 $\xi_k = \mathbf{0}$,则停止,否则转(3).

(3) 在 S 上求解问题(P_3),得最优解,记为(x_{k+1}, v_{k+1}),令 $k = k+1$,返回(2).

算法3在每次迭代中增加一个约束,从几何意义上看,就是用一个超平面将 S 中不包含解的部分割掉.这在实现上将无限制地增加约束,因而计算量很大.

定理3 设 $f(x)$ 是凸的且下方有限,则算法3产生的点列 $\{(x_{k+1}, v_{k+1})\}$ 满足

1° $v_2 \leqslant v_3 \leqslant \cdots \leqslant v_k \to f^* = \min\limits_{x \in S} f(x')$.

2° $\{x_k\}$ 的任一聚点都是 $f(x)$ 在 S 上的极小点.

5.2.3 束法

束法(bundle method)或称束方法,是从共轭次梯度法发展而得到的一种方法,它是一种下降算法,要求在每一迭代点有 $f(x_{k+1}) \leqslant f(x_k)$.

首先介绍共轭次梯度法.在第 k 次迭代时,有一个指标集 $I_k \subset \{1,2,\cdots,k\}$,其搜索方向由

$$d_k = -\sum_{i \in I_k} \lambda_i^{(k)} \xi_i, \quad \xi_i \in \partial f(x_k) \tag{5-12}$$

给出,其中 $\lambda_i^{(k)}, i \in I_k$ 是通过求解子问题

(P_4)
$$\min \| \sum_{i \in I_k} \lambda_i \xi_i \|^2,$$
$$\text{s.t.} \sum_{i \in I_k} \lambda_i = 1, \lambda_i \geqslant 0, i \in I_k$$

得到的.当 $f(x)$ 是凸的二次函数以及 $I_k = \{1,2,\cdots,k\}$ 时,则在精确线搜索下,由问题(P_4)所产生的方向和共轭次梯度法是一致的.下面给出共轭次梯度法的具体步骤.

算法4

(1) 给定初始点 $x_0 \in \mathbf{R}^n$, $\xi_0 \in \partial f(x_0)$.选取 $0 < m_2 < m_1 < \frac{1}{2}$, $0 < m_3 < 1$, $\epsilon > 0$, $\eta > 0$,令 $I_0 = \{0\}$, $k = 0$.

(2) 求解问题(P_4),确定 $\lambda_i^{(k)}$ 后代入(5-12)式得 d_k,如果 $\| d_k \| \leqslant \eta$,则停止,否则转(3).

(3) 计算 $y_k = x_k + \alpha_k d_k$,使得

$$f(y_k) \leqslant f(x_k) - m_2 \alpha_k \| d_k \|^2$$

或者

$$\| y_k - x_k \| \leqslant m_3 \epsilon$$

成立.

(4) 如果存在 $\xi_{k+1} \in \partial f(y_{k+1})$,使得

$$\xi_{k+1}^{\mathrm{T}} d_k \geqslant - m_1 \| d_k \|^2,$$

则令 $x_{k+1}=y_k$，否则置 $x_{k+1}=x_k$.

(5) 令 $I_{k+1}=I_k\cup\{k+1\}\setminus T_k$，其中 T_k 是所有满足 $\|x_i-x_{k+1}\|>\epsilon$ 的下标 i 的集合.

(6) 令 $k=k+1$，返回(2).

定理 4 设 $f(x)$ 是凸的，$\|\partial f(x)\|$ 在包含集合 $\{x\in\mathbf{R}^n \mid f(x)\leqslant f(x_0)\}$ 的某一开集上是有界的，由算法 4 产生的点列 $\{x_k\}$ 使得 $f(x_k)$ 下方有界，则算法 4 必经过有限次迭代后终止.

下面讨论共轭次梯度法的推广.

在第 k 次迭代时，有加权因子 $t_i^{(k)}\geqslant 0,\ i=0,1,2,\cdots,k$. 考虑子问题

$$
(\mathrm{P}_5)\qquad \begin{aligned} &\min\Big\|\sum_{i=1}^{k}\lambda_i\xi_i\Big\|^2,\\ &\text{s.t.}\quad \sum_{i=1}^{k}\lambda_i=1,\ \lambda_i\geqslant 0,\\ &\qquad\ \ \sum_{i=1}^{k}\lambda_i t_i^{(k)}\leqslant\bar{\epsilon},\end{aligned}
$$

其中 $\bar{\epsilon}$ 是预先给定的常数. 记 $\lambda_i^{(k)}$ 是问题(P_5)的解，束法的搜索方向取为

$$
d_k=-\sum_{i=1}^{k}\lambda_i^{(k)}\xi_i. \tag{5-13}
$$

不难看出，若取 $t_i^{(k)}=0,\ i\in I_k$ 和 $t_i^{(k)}=+\infty,\ i\bar{\in} I_k$，则问题($\mathrm{P}_4$)和($\mathrm{P}_5$)等价. 下面给出束法的具体步骤.

算法 5

(1) 给定初始点 $x_0\in\mathbf{R}^n$，$\xi_0\in\partial f(x_0)$，选取 $0<m_2<m_1<\dfrac{1}{2}$，$0<m_3<1$，$\epsilon>0$，$\eta>0$，令 $t_0^{(0)}=1$，$k=0$.

(2) 求解问题(P_5)，确定 $\lambda_i^{(k)}$ 后代入(5-13)式得 d_k，如果 $\|d_k\|\leqslant\eta$，则停止，否则转(3).

(3) 计算 $y_k=x_k+\alpha_k d_k$ 使得

$$
f(y_k)\leqslant f(x_k)-m_2\alpha_k\|d_k\|^2 \tag{5-14}
$$

或者

$$
f(y_k)-\alpha_k\xi_{k+1}^{\mathrm{T}}d_k\geqslant f(x_k)-\epsilon \tag{5-15}
$$

成立，其中 $\xi_{k+1}\in\partial f(y_k)$. 若(5-14)式不成立，则转(5).

(4) 令 $x_{k+1}=y_k$，$t_{k+1}^{(k+1)}=1$，

$$
t_j^{(k+1)}=t_j^{(k)}+f(x_{k+1})-f(x_k)-\alpha_k\xi_j^{\mathrm{T}}d_k,\quad j=0,1,2,\cdots,k,
$$

$k=k+1$，返回(2).

(5) 令 $x_{k+1}=x_k$，$t_j^{(k+1)}=t_j^{(k)}$，$j=0,1,2,\cdots,k$，

$$
t_{k+1}^{(k+1)}=f(x_k)-f(y_k)+\alpha_k\xi_{k+1}^{\mathrm{T}}d_k,
$$

$k=k+1$，返回(2).

算法 5 与算法 4 具有相同的收敛性结果.

5.3 利普希茨规划算法

本节假定问题(P_1)中的 $f(x)$ 是局部利普希茨函数.下面将给出求解(P_1)的束方法.

假定 $\{x_k\}$ 是算法产生的点列,d_k 是搜索方向,α_k^L 是步长,$x_{k+1} = x_k + \alpha_k^L d_k$.引入辅助点列 $\{y_k\}$,其中 $y_{k+1} = x_k + \alpha_k^R d_k$,$\alpha_k^R$ 是辅助步长,满足 $\alpha_k^L \leqslant \alpha_k^R$.给定 $y \in \mathbf{R}^n$ 及 $\partial f(y)$ 中一个元素 $\xi(y)$,定义 $f(x)$ 一个线性化为

$$\bar{f}(x;y) = f(y) + \langle \xi(y), x - y \rangle, \forall x \in \mathbf{R}^n,$$

线性化误差记为

$$\alpha(x,y) = \max\{f(x) - \bar{f}(x;y), r\|x - y\|^2\},$$

其中 $r \geqslant 0$ 为给定常数,当 $f(x)$ 是凸函数时,取 $r = 0$.在第 k 步迭代时,有指标集 $I_k \subset \{1,2,\cdots,k\}$ 和线性化函数 $f_j(\cdot) = f(\cdot;y_j)$,$j \in I_k$,其中

$$f_j(x) = f_j^k + \langle \xi(y_j), x - x_k \rangle, \forall x \in \mathbf{R}^n,$$

$f_j^k = \bar{f}(x_k;y_j)$,$j \in I_k$.下面计算中用下式来近似 $\alpha(x_k,y_j)$:

$$\alpha_j^k = \max\{f(x_k) - f_j^k, r(s_j^k)^2\},$$

其中 $s_j^k = |y_j - x_j| + \sum_{i=j}^{k-1} |x_{i+1} - x_i|$.下面给出算法具体步骤.

算法 6

(1) 给定初始点 $x_1 \in \mathbf{R}^n$,$\xi_1 \in \partial f(x_1)$.选取参数 $\eta \geqslant 0$,$e_a > 0$,m_L,m_R,m_e,$\bar{\alpha}$,$\tilde{\alpha}$ 满足 $m_L < m_R < 1$,$m_e < 1$,$\bar{\alpha} \leqslant 1 \leqslant \tilde{\alpha}$;选取 $r > 0$(当 $f(x)$ 是凸函数时,取 $r = 0$),$\bar{a} > 0$,置 $I_1 = \{1\}$;$y_1 = x_1$,$p_0 = \xi_1$,$f_p^1 = f_1^1 = f(y_1)$,$s_p^1 = s_1^1 = 0$,$e^1 = e_a$,$a^1 = 0$,$r_a^1 = 1$;令 $k = 1$,$l = 0$,$k(0) = 1$.

(2) 解下述问题确定 $\lambda_j^k (j \in I_k)$,λ_p^k.

$$\min \frac{1}{2} \left| \sum_{j \in I_k} \lambda_j g_j + \lambda_p p_{k-1} \right|^2,$$

$$\text{s.t.} \quad \lambda_j \geqslant 0, j \in I_k, \lambda_p \geqslant 0, \quad \sum_{j \in I_k} \lambda_j + \lambda_p = 1,$$

$$\lambda_p = 0(\text{如果 } \gamma_a^k = 1),$$

$$\sum_{j \in I_k} \lambda_j \alpha_j^k + \lambda_p \alpha_p^k \leqslant e^k.$$

置

$$(p_k, \tilde{f}_p^k, \tilde{s}_p^k) = \sum_{j \in I_k} \lambda_j^k (\xi_j, f_j^k, s_j^k) + \lambda_p^k (p^{k-1}, f_p^k, s_p^k),$$

$$d_k = -p_k, v_k = -\|p_k\|^2.$$

若 $\lambda_p^R = 0$,置

$$a^k = \max\{s_j^k \mid j \in I_k\}.$$

(3) 置 $\tilde{\alpha}_p^k = \max\{|f(x_k) - \tilde{f}_p^k|, r(\tilde{s}_p^k)^2\}$,若 $\max\{\|p_k\|^2, \tilde{\alpha}_p^k\} \leqslant \eta$,则停止,否则转(4).

(4) 若 $\|p_k\|^2 > \tilde{\alpha}_p^k$,转(5),否则用 $m_e e^k$ 代替 e^k,返回(2).

(5) 利用线搜索来确定步长 α_k^L 和 α_k^R,满足 $0 \leqslant \alpha_k^L \leqslant \alpha_k^R \leqslant \tilde{\alpha}$,当 $\alpha_k^L > 0$ 时,$\alpha_k^R = \alpha_k^L$、$x_{k+1} = x_k + \alpha_k^L d_k$ 和 $y_{k+1} = x_k + \alpha_k^R d_k$ 满足

$$f(x_{k+1}) \leqslant f(x_k) + m_L \alpha_k^L v_k;$$

$\alpha_k^L \geqslant \bar{\alpha}$ 或当 $\alpha_k^L > 0$ 时,$\alpha(x_k, x_{k+1}) > m_e e^k$;

当 $\alpha_k^L = 0$ 时,$\alpha(x_k, y_{k+1}) \leqslant m_e e^k$;

当 $\alpha_k^L = 0$,$\langle \xi(y_{k+1}), d_k \rangle \geqslant m_R v_k$.

(6) 若 $\alpha_k^L = 0$,置 $e^{k+1} = e^k$,否则,选取 $e^{k+1} > e_a$,令

$$k(l+1) = k+1, l = l+1.$$

(7) 选取 $\hat{I}_k \subset I_k$ 使得 $I_{k+1} = \hat{I}_k \cup \{k+1\}$ 包含 $k(l)$.选取 $\xi_{k+1} \in \partial f(y_{k+1})$,计算

$$f_{k+1}^{k+1} = f(y_{k+1}) + \langle \xi_{k+1}, x_{k+1} - y_{k+1} \rangle,$$

$$f_j^{k+1} = f_j^k + \langle \xi_j, x_{k+1} - x_k \rangle, j \in \hat{I}_k,$$

$$f_p^{k+1} = \tilde{f}_p^k + \langle p_k, x_{k+1} - x_k \rangle,$$

$$s_{k+1}^{k+1} = \|y_{k+1} - x_{k+1}\|,$$

$$s_j^{k+1} = s_j^k + \|x_{k+1} - x_k\|, j \in \hat{I}_k,$$

$$s_p^{k+1} = \tilde{s}_p^k + \|x_{k+1} - x_k\|.$$

(8) 置 $a_{k+1} = \max\{a_k + \|x_{k+1} - x_k\|, s_{k+1}^{k+1}\}$,若 $a^{k+1} \leqslant \bar{a}$ 或 $\alpha_k^L = 0$,置 $r_a^{k+1} = 0$,转(10),否则置 $r_a^{k+1} = 1$,转(9).

(9) 在 I_{k+1} 中删除满足 $s_j^{k+1} > \bar{a}/2$ 的指标 j,置 $a^{k+1} = \max\{s_j^{k+1} \mid j \in I_{k+1}\}$.

(10) 令 $k = k+1$,返回(2).

定理 5 设 $\{x_k\}$ 是由算法 6 产生的点列,若有 $x_k \to \bar{x}$,$\|p_k\| \to 0$,$\tilde{\alpha}_p^k \to 0(k \in K)$,其中 $K \subset \{1,2,\cdots\}$,则有 $0 \in \partial f(\bar{x})$.

5.4 一类复合不可微优化算法

本节讨论具有如下特殊形式的不可微优化问题:

(P$_6$)
$$\min_{x \in \mathbb{R}^n} h(f(x)).$$

其中 $f(x) = (f_1(x), f_2(x), \cdots, f_m(x))^{\mathrm{T}}$ 是连续可微的,$h(\cdot)$ 是凸函数.问题(P$_6$)

的目标函数是复合函数，故称**复合不可微优化**. 复合不可微优化具有很强的实际背景，例如第 1 章(1-1) 式、(1-4) 式和(1-7) 式所给出的函数，都属复合不可微函数.

引入下述记号：

$$\chi(x;d) = h(f(x)) - h(f(x) + A(x)^{\mathrm{T}}d),$$
$$\psi_t(x) = \max_{\|d\| \leqslant t} \chi(x;d),$$
$$\mathrm{D}F(x;d) = \sup_{\lambda \in \partial h(f)|_{f=f(x)}} d^{\mathrm{T}}A(x)\lambda,$$

其中 $\partial h(f)|_{f=f(x)}$ 是指函数 $h(\cdot)$ 在 $f(x)$ 处的次梯度，$A(x) = \nabla f(x)^{\mathrm{T}}$.

利用复合函数广义梯度链锁规则不难验证，$0 \in \partial(h(f(x)))$ 等价于 $\mathrm{D}F(x,d) \geqslant 0, \forall d \in \mathbf{R}^n$.

对于复合不可微优化问题(P_6)，**信赖域法**的子问题具有下述形式：

$$(\mathrm{P}_7) \qquad \min h(f(x_k)) + A(x_k)^{\mathrm{T}}d + \frac{1}{2}d^{\mathrm{T}}B_k d = \phi_k(d),$$
$$\text{s.t.} \quad \|d\| \leqslant \Delta_k,$$

其中 $A(x) = \nabla f(x)$，B_k 是一 $n \times n$ 阶对称阵.

下面给出求解问题(P_6) 的**信赖域法**.

算法 7

(1) 给定初始点 $x_1 \in \mathbf{R}^n$，及 $\lambda_0 \in \mathbf{R}^m$，$\Delta_1 > 0$，$\eta \geqslant 0$，令 $k = 1$.

(2) 计算 $B_k = \sum_{i=1}^{m}(\lambda_{k-1})_i \nabla^2 f_i(x_k)$，求解子问题($\mathrm{P}_7$)，记($\mathrm{P}_7$) 的最优解为 d_k，如果 $\|d_k\| \leqslant \eta$，则停止，否则转(3).

(3) 计算

$$\gamma_k = \frac{h(f(x_k)) - h(f(x_k + d_k))}{\phi_k(0) - \phi_k(d_k)},$$

如果 $\gamma_k < 0.25$，则令 $\Delta_{k+1} = \|d_k\|/4$；如果 $\gamma_k > 0.75$ 且 $\|d_k\| = \Delta_k$，则令 $\Delta_{k+1} = 2\Delta_k$；如果 Δ_{k+1} 还未定义，则令 $\Delta_{k+1} = \Delta_k$.

(4) 如果 $\gamma_k > 0$，则转(5)；否则令 $x_{k+1} = x_k$，$\lambda_k = \lambda_{k-1}$，转(6).

(5) $x_{k+1} = x_k + d_k$；λ_k 由下述方程确定：

$$A(x_k)\lambda_k + B_k d_k + \overline{\mu}_k \mu_k = 0,$$
$$\overline{\mu}_k(\Delta_k - \|d_k\|) = 0,$$

其中 $\lambda_k \in \partial h(\cdot)|_{f(x_k)+A(x_k)^{\mathrm{T}}d_k}$，$\mu_k \in \partial\|\cdot\|_{d_k}$，$\overline{\mu}_k \geqslant 0$.

(6) 令 $k = k + 1$，返回(2).

5.5 非光滑方程组的解法

变分、互补以及非线性规划问题，都可等价地转化为求解一个非光滑(不可微)方程组.

设 $F:\mathbf{R}^n \to \mathbf{R}^m$ 是局部利普希茨向量函数，$F(x)$ 在点 $x \in \mathbf{R}^n$ 处称为是半光滑

的,如果对任意 $h \in \mathbf{R}^n$,有

$$\lim_{\substack{V \in \partial F(x+th') \\ h' \to h, t \to 0^+}} Vh'$$

存在,其中 $\partial F(\cdot)$ 代表 $F(\cdot)$ 的克拉克广义雅可比.

半光滑函数类包括许多常见的非光滑函数,例如,凸函数、光滑函数的极大值函数等.同时,半光滑函数的四则运算以及光滑复合等均是半光滑函数.目前人们所讨论的非光滑方程组基本上限于半光滑函数.

考虑下述**非光滑方程组**

$$F(x) = 0, \tag{5-16}$$

其中 $F:\mathbf{R}^n \to \mathbf{R}^m$ 是局部利普希茨函数且为半光滑的.求解方程(5-16)的**广义牛顿法**的迭代公式如下:

$$x_{k+1} = x_k - V_k^{-1}F(x_k), V_k \in \partial F(x_k). \tag{5-17}$$

命题 2 设 $F:\mathbf{R}^n \to \mathbf{R}^m$ 是局部利普希茨函数,对任意 $V \in \partial F(x)$,矩阵 V 是非奇异的,那么存在 x 的一个邻域 $N(x)$ 和常数 C,使得对任意 $y \in N(x)$,$V \in \partial F(y)$ 是非奇异的,且有 $\| V^{-1} \| \leqslant C$.

定理 6 假设 $x^* \in \mathbf{R}^n$ 是方程(5-16)的解,$F(x)$ 是 $\mathbf{R}^n$ 上的局部利普希茨向量函数,在 x^* 处是半光滑的,如果任意 $V \in \partial F(x^*)$ 是非奇异的,那么迭代公式(5-17)是适定的,且当初始点与 x^* 充分近时,由公式(5-17)产生的点列 $\{x_k\}$ 收敛于 x^*.

证明 由命题 2 知,迭代公式(5-17)是适定的.通过计算得

$$\begin{aligned}
\| x_{k+1} - x^* \| &= \| x_k - x^* - V_k^{-1}F(x_k) \| \\
&\leqslant \| V_k^{-1}(F(x_k) - F(x^*) - F'(x^*;x_k - x^*)) \| + \\
&\quad \| V_k^{-1}(V_k(x_k - x^*) - F'(x^*;x_k - x^*)) \| \\
&= o(\| x_k - x^* \|).
\end{aligned}$$

在上式推导中,利用了命题 2 中结论 $\| V^{-1} \| \leqslant C$ 和半光滑函数的性质:$F'(x^*;\cdot)$ 存在,且有 $Vh - F'(x^*;h) = o(\| h \|)$,$\forall V \in \partial F(x^* + h)$ 以及 $F(x^* + h) - F(x^*) - F'(x^*;h) = o(\| h \|)$.定理得证.

利用迭代公式(5-17)求解方程(5-16),需要在每一迭代点 x_k 处能够计算克拉克广义雅可比 $\partial F(x_k)$ 中的一个元素,这对于一般的非光滑函数是做不到的.但是,对于较特殊的非光滑函数,例如,由非线性互补问题转化的非光滑函数,这是能够办到的.

参 考 文 献

1 Clarke F H. Optimization and nonsmooth analysis. New York: Wiley-Interscience, 1983.

2 Demyanov V F, Rubinov A M. Constructive nonsmooth analysis. Berlin: Peterlang, 1995.

3 Demyanov V F, Vasiliev L V. Nondifferentiable optimization. New York: Optimization Software, 1986.

4　Hiriart-Urruty J B, Lemarechal C. Convex analysis and minimization algorithm. Berlin: Springer, 1993.

5　Makela M M, Neittaanmaki P. Nonsmooth optimization——analysis and algorithms with applications to optimal control. Singapore: World Scientific, 1992.

第9篇

整数规划

编　者　施光燕
审校者　马仲蕃

目　录

引　言

整数规划是要求变量取整数值的数学规划.要求变量取整数的线性规划,称为**线性整数规划**.变量只取0或1的规划称为**0-1规划**.只要求部分变量取整数值的规划,称为**混合整数规划**.由于实际中有许多量必须是整数,如人数、机器数、运输次数等;又由于利用0,1变量可以数量化地描述开与关、取与弃、有与无等逻辑现象,所以在很多领域中,如线路设计、工厂选址、人员安排、课程安排、代码选取等,常常出现整数规划问题.自1959年柯莫瑞(R.E.Gomory)提出了解线性整数规划的割平面法后,整数规划就逐步形成为一个独立的分支.

1　整数规划模型举例

1.背包问题

一个背包的容积为v,现有n种物品可装,物品j的质量为W_j,体积为$v_j(j=1,2,\cdots,n)$.问如何配装才能既不超过背包的容积,又能使装的总质量最大.

设变量

$$x_j=\begin{cases}1, & \text{物品 } j \text{ 被装入背包;}\\ 0, & \text{物品 } j \text{ 不装入背包,}\end{cases}$$

则问题可写成如下的数学规划形式:

求　　$$\max\sum_{j=1}^{n}W_jx_j$$

满足$\sum\limits_{j=1}^{n}v_jx_j\leqslant v,\ x_j=0$ 或 $1\quad j=1,2,\cdots,n.$

2.工厂选址问题

有n个城市$\{1,2,\cdots,n\}$,每日需要某种物资的数量,分别为$d_1,d_2,\cdots,d_n$.现在计划要在其中选取m个城市,建造m座生产这种物资的工厂.若在城市j处建厂,假设已知日产量最多只能为S_j,而生产投资为F_j,设从城市i到城市j的单位运价为c_{ij},试问m个工厂应设在何处,才能既满足需要又使总投资最省.

设变量

$$y_i=\begin{cases}1, & \text{在城市 } i \text{ 中建厂;}\\ 0, & \text{不在城市 } i \text{ 中建厂.}\end{cases}$$

设x_{ij}为从城市i运给城市j的物资数量,则问题可写成如下的规划形式:

求　　$$\min\{\sum_i\sum_j c_{ij}x_{ij}+\sum_i F_iy_i\}$$

满足
$$\sum_{j=1}^{n} x_{ij} \leqslant S_i y_i, \quad i = 1,2,\cdots,n;$$
$$\sum_{i=1}^{n} x_{ij} \geqslant d_j, \quad j = 1,2,\cdots,n;$$
$$\sum_{i=1}^{n} y_i = m, y_i = 0 \text{ 或 } 1; x_{ij} \geqslant 0, i,j = 1,2,\cdots,n.$$

3. 加工问题

有 m 台同一类型的机床，有 n 种零件要在这种机床上加工. 设每种零件所需的加工时间分别为 $a_1, a_2, \cdots, a_n$，试问如何分配才能使各机床的总加工任务相等，或者说，尽可能均衡.

设变量

$$x_{ij} = \begin{cases} 1, & \text{若 } a_j \text{ 分配在机床 } i \text{ 上加工;} \\ 0, & \text{若 } a_j \text{ 不在机床 } i \text{ 上加工.} \end{cases}$$

则问题可写成如下的形式:

求
$$\min\left\{\max\left[\sum_{j=1}^{n} x_{1j}a_j, \sum_{j=1}^{n} x_{2j}a_j, \cdots, \sum_{j=1}^{n} x_{mj}a_j\right]\right\},$$

满足
$$\sum_{i=1}^{m} x_{ij} = 1, j = 1,2,\cdots,n,$$
$$x_{ij} = 0 \text{ 或 } 1.$$

4. 系统可靠性问题

有 n 个元件组成一个并联(或串联)系统. 设第 i 个位置所用的元件可从集合 S_i 中挑选. 设 c_{ij} 为元件 j 用在第 i 个位置上所花的费用，$j \in S_i$，而 P_{ij} 表示其可靠性概率. 试问应如何选择各位置的元件使得系统的可靠性概率 $P \geqslant a$，且使系统总的费用最省.

设

$$x_{ij} = \begin{cases} 1, & \text{若 } j \in S_i\text{，且元件 } j \text{ 用在位置 } i \text{ 上;} \\ 0, & \text{若 } j \in S_i\text{，但元件 } j \text{ 不用在位置 } i \text{ 上,} \end{cases}$$

则并联系统的可靠性概率为

$$P = 1 - \prod_{i=1}^{n}\left[\prod_{j \in S_i}(1 - P_{ij})^{x_{ij}}\right],$$

条件 $P \geqslant a$ 可写为

$$1 - a \geqslant \prod_{i=1}^{n}\left[\prod_{j \in S_i}(1 - P_{ij})^{x_{ij}}\right],$$

即
$$\log(1 - a) \geqslant \sum_{i=1}^{n}\sum_{j \in S_i} x_{ij}\log(1 - P_{ij}).$$

设
$$a_{ij} = \log(1 - P_{ij}), k = \log(1 - a),$$

则问题(并联情形)可写成如下形式

求
$$\min\sum_{i=1}^{n}\sum_{j\in S_i}c_{ij}x_{ij},$$
满足
$$\sum_{i=1}^{n}\sum_{j\in S_i}a_{ij}x_{ij}\leqslant k,$$
$$\sum_{j\in S_i}x_{ij}=1,\quad i=1,2,\cdots,n,$$
$$x_{ij}=0\text{ 或 }1,\quad i=1,2,\cdots,n,\quad j\in S_i$$
(对串联情形,其中的 $a_{ij}=-\log P_{ij}, K=-\log a$).

5. 离散值的变量

设变量 x_j 只能取 k 个数值 $\{a_1,a_2,\cdots,a_k\}$ 中的一个(例如材料的规格),则 x_j 可表示为
$$x_j=\sum_{i=1}^{k}a_iy_i,\quad \sum_{i=1}^{k}y_i=1,\quad y_i=0\text{ 或 }1.$$

6. 跳跃变量

设变量 x_j 的值,或者为 0 或者满足
$$L\leqslant x_j\leqslant U,$$
则可用下述条件表示跳跃变量:
$$x_j\geqslant Ly,\ x_j\leqslant Uy,\ y=0\text{ 或 }1.$$

例 1　最轻重量结构设计　平面门式框架受 3 种载荷作用,结构的几何尺寸如图 1-1 所示. 弹性模量 $E=206.88\text{GPa}$,许用正应力 $[\sigma]=163.86\text{MPa}$, 材料容重 $\rho=76\,999.34\text{N/m}^3$,节点水平位移的上限 $\bar{\delta}=12.7\text{mm}$.

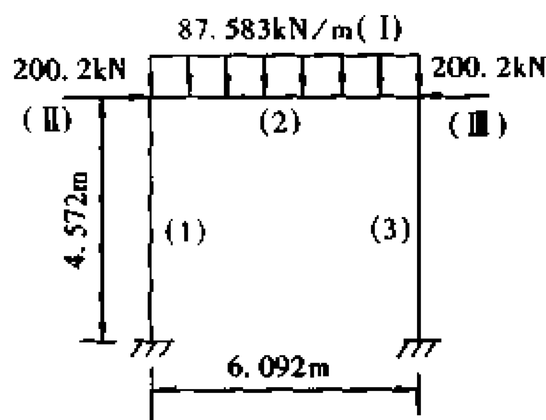


图 1-1

结构的设计变量是各截面的几何性质,其可供选择的变量为下面的离散变量集.

序号	1	2	3	4	5	6
A/cm^2	99.00	105.83	112.26	134.92	140.01	144.99
W/cm^3	1293.46	1429.71	1561.75	2057.72	2175.33	2290.86
I/cm^4	29136.20	33298.50	37460.30	54110.10	58272.40	62434.70

设变量
$$x_{ik}=\begin{cases}1, & \text{第 } i \text{ 号杆采用第 } k \text{ 号几何性质;}\\ 0, & \text{第 } i \text{ 号杆不采用第 } k \text{ 号几何性质,}\end{cases}$$
$i=1,2,3; k=1,2,3,4,5,6.$ 问题归结为

求
$$\min\sum_{i=1}^{3}\rho_il_i\Big(\sum_{k=1}^{6}A_kx_{ik}\Big),$$

满足

$$\frac{|N_{i1}|}{\sum\limits_{k=1}^{6}A_kx_{ik}}+\frac{|M_{i1}|}{\sum\limits_{k=1}^{6}W_kx_{ik}}\leqslant 163.86\times10^6\ \mathrm{Pa},$$

$$\sum_{i=1}^{3}\left(\frac{N_iN_{i1}l_i}{E\sum\limits_{k=1}^{6}A_kx_{ik}}+\int_{l_i}\frac{M_iM_{i1}}{E\sum\limits_{k=1}^{6}I_kx_{ik}}\mathrm{d}x\right)\leqslant 12.7\times10^{-3}\mathrm{m},$$

$$x_{ik}=0\ 或\ 1,\quad i=1,2,3,\quad k=1,2,3,4,5,6.$$

2　线性整数规划基本解法

考虑线性整数规划

(P)
$$\max z=\sum_{i=1}^{n}c_ix_i,$$

$$\text{s.t.}\sum_{j=1}^{n}a_{ij}x_j=b_i,\ i=1,2,\cdots,m,$$

2.1　基本解法概述

关于整数规划,有如下两种基本的想法:

(1) 因为可行方案的数目常常是有限的,因而通过一一比较后,总能求得最优方案.例如,背包问题的装法,充其量有 2^{n-1} 种方式.但这种穷举的办法,实际上是不可行的,因若设想计算机每秒钟能比较100万个方式,要比较完 2^{45} 种方式,则要用一年多时间.

(2) 先放弃变量取整的要求,把一个线性整数规划化为一个线性规划,然后求得线性规划解后采用"四舍五入"的方法求整数解.这种方法只有在变量的取值较大时,才有成功的可能性,而当变量的取值较小时,特别是0-1变量时,往往不会成功.例如,整数规划问题:

$$\max z=3x_1+13x_2,$$

$$\text{s.t.}\quad 2x_1+9x_2\leqslant 40,\ 11x_1-8x_2\leqslant 82,$$

$$x_1,x_2\geqslant 0\ 且取整数值.$$

放弃整数性条件后线性规划的解为 $x_1^*=9.2,x_2^*=2.4$,而 $z^*=58.8$,其附近的4个整数点(9,2),(10,2),(10,3),(9,3)均不是整数规划的可行解.原整数规划的最优解为 $x_1=2,x_2=4,z=58$.

然而,若能给出线性整数规划所有可行点的最小凸包(整点凸包),则在此凸包上求线性规划的最优解,便可得到线性整数规划的最优解.但是求整点凸包是整数规划中一个很难的基本理论问题,目前尚未解决.在研究凸包过程中,柯莫瑞首先

提出割平面的概念，并建立了整数规划的基本算法 —— 割平面法.

分支定界法是目前解整数规划最实用的算法之一. 它用到整数规划解法中三个基本概念.

1. 分解

对任何整数规划问题(P)，让 $F(P)$ 表示(P) 的可行解集合，问题(P) 称为**分解**成子问题$(P_1),(P_2),\cdots,(P_k)$ 之和，若满足条件

$$\bigcup_{i=1}^{k} F(P_i) = F(P),\quad F(P_i)\cap F(P_j) = \varnothing, 1 \leqslant i \neq j \leqslant k.$$

通常分解的方式是“两分法”，即若 x_j 是整变量，则把问题(P) 按照条件 $x_j \leqslant 4$ 和 $x_j \geqslant 5$ 分解成两个子问题之和.

2. 松弛

对任何整数规划问题(P)，凡是放弃(P) 的某些约束条件后所得到的问题($\tilde{\mathrm{P}}$)，都称为(P) 的**松弛问题**. 对于(P) 的任何松弛问题($\tilde{\mathrm{P}}$)，都具有如下明显的性质：

(1) 若($\tilde{\mathrm{P}}$ 没有可行解，则(P) 也没有可行解；

(2) 对求最大值的目标函数而言，(P) 的最大值不大于($\tilde{\mathrm{P}}$)的最大值；

(3) 若($\tilde{\mathrm{P}}$) 的最优解是(P) 的可行解，则($\tilde{\mathrm{P}}$) 的最优解就是(P) 的一个最优解.

最通常的松弛方法是放弃变量的整数性要求.

3. 探测

假设按某种规则已将问题(P) 分解成子问题$(P_1),(P_2),\cdots,(P_k)$ 之和，并且各(P_i) 已有对应的松弛问题($\tilde{\mathrm{P}}_i$).

(F1) 若($\tilde{\mathrm{P}}_i$) 没有可行解，则探明了(P_i) 没有可行解. 因此，可从(P) 的分解表上把它删去.

(F2) 假设已掌握了(P) 的一个可行解 $\bar{x}$，它的目标函数值为 $\bar{z}$，若松弛问题($\tilde{\mathrm{P}}_i$) 的最大值(针对求 max 而言) 不比 $\bar{z}$ 大，则探明(P_i) 中没有比 $\bar{x}$ 更好的可行解. 因此无须再考虑(P_i)，可从分解表上把它删去.

(F3) 若($\tilde{\mathrm{P}}_i$) 的最优解是(P_i) 的可行解，则已求得了(P_i) 的一个最优解. 因此也无须进一步考虑(P_i) 了，可从分解表上把它删去. 同时，若(P_i) 的最优解比 $\bar{x}$ 好，则替换 $\bar{x}$ 并相应替换 $\bar{z}$.

(F4) 假如表上各个($\tilde{\mathrm{P}}_i$) 的目标函数最大值都不比 $\bar{x}$ 大，那么，当时的记录解 $\bar{x}$，便是原问题(P) 的一个最优解.

通常，称情形(F1) 为**可行性探测**，称(F2)，(F3)，(F4) 为**最优性探测**.

结论 求解整数规划问题(P)，可归纳为

选定一种松弛方式，将(P) 松弛成问题($\tilde{\mathrm{P}}$)，使得较易求解. 若($\tilde{\mathrm{P}}$) 无可行解，则(P) 也无可行解. 若($\tilde{\mathrm{P}}$) 的最优解是(P) 的可行解，则已求得(P) 的最优解.

若($\tilde{\mathrm{P}}$)的最优解不是(P)的可行解,则有两条不同的途径可走:一条途径是设法改进松弛问题($\tilde{\mathrm{P}}$),坚持探测(P)(例如采用割平面法);另一条路径是选定一种分解方式,把(P)分解成两个或几个子问题之和,将其列表记录下来,并且赋予各子问题一个尽可能好的目标函数值的上界,然后按一定次序,逐个进行探测.当某个子问题已经被探明时,就从表中删去;否则继续对子问题进行分解(例如采用分支定界法).

2.2 分支定界法

线性整数规划问题(P),其分支定界法的步骤如下:

(1)取($\tilde{\mathrm{P}}$)为放弃 x_j 的整数性要求后的线性规划问题.用单纯形法求解($\tilde{\mathrm{P}}$).

若($\tilde{\mathrm{P}}$)无解,则终止,(P)亦无解;

若($\tilde{\mathrm{P}}$)的最优解 $\bar{x}$ 为整数解,则终止,$\bar{x}$ 即为(P)的最优解;

若($\tilde{\mathrm{P}}$)的最优解 $\bar{x}$ 不是整数解,则取相应的目标函数值作为(P)的最优值的上界 $\bar{z}$.转(2).

(2)置分解表 $\Pi=\{(\mathrm{P})\}$,$x^*=\varnothing$,$z^*=-\infty$.

(3)若 $\Pi=\varnothing$,则终止,x^* 便是最优解,z^* 即为最优值;否则转(4).

(4)从 Π 中任选一子问题(CP)(可选上界值最大的子问题,或以先入后出的原则选).$\Pi/(\mathrm{CP})\to\Pi$.

(5)对(CP)放弃 x_j 的整数性要求得$(\widetilde{\mathrm{CP}})$,用单纯形法解$(\widetilde{\mathrm{CP}})$.

若$(\widetilde{\mathrm{CP}})$无可行解或最大值 $\tilde{z}\leqslant z^*$,则返回(3);

若$(\widetilde{\mathrm{CP}})$的最优解 $\tilde{x}$ 为非整数解,则转(6);

否则,$\tilde{z}\to z^*$,$\tilde{x}\to x^*$ 返回(3).

(6)从 $\tilde{x}$ 中任取一分数分量 $\tilde{x}_j$,按条件 $x_j\leqslant[\tilde{x}_j]$ 和 $x_j\geqslant[\tilde{x}_j]+1$ 将(CP)分解为两个子问题(CP_1)和(CP_2),并赋予它们的上界为$[\tilde{z}]$(其中符号$[a]$表示不超过 a 的最大整数).将(CP_1)和(CP_2)以及它们的上界记入 Π 中.返回(4).

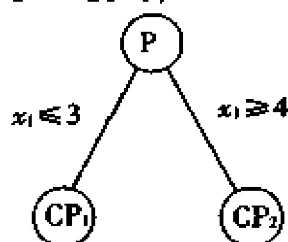


图 2-1

例1 求解整数规划:

$$(\mathrm{P})\quad \max z=-(x_1+4x_2),$$
$$\text{s.t.}\quad 2x_1+x_2\leqslant 8,\ x_1+2x_2\geqslant 6,$$
$$x_1,x_2\geqslant 0$$

且为整数.

解 放弃整数约束,作为普通线性规划求解,得到最优解:$x_1=10/3$,$x_2=4/3$,最优值 $z=-26/3$.由算法有 $\Pi=\{(\mathrm{P})\}$,$\bar{z}=-26/3$,$x^*=\varnothing$,$z^*=-\infty$.

因为 x_1, x_2 均为非整数，设选 x_1 进行分支．分支过程的树形图如图2-1所示．按 $x_1 \leqslant [10/3] = 3$ 及 $x_1 \geqslant [10/3] + 1 = 4$ 分成两个子问题 CP_1, CP_2．其中问题(CP_1)为在问题(P)的基础上增加约束 $x_1 \leqslant 3$ 而得，问题(CP_2)为在问题(P)的基础上增加约束 $x_1 \geqslant 4$ 而得，从树形图2-1上由根部至节点支上的说明即可知，所以一般不标 P, CP_1, CP_2，而标以求解的序号．

设下步放弃(CP_1)中整数要求，解相应的线性规划，得最优解为 $x_1 = 3, x_2 = 3/2, z = -9$．由算法再根据 $x_2 = 2/3$，按 $x_2 \leqslant [3/2] = 1$ 和 $x_2 \geqslant [3/2] + 1 = 2$ 分支，如图2-2所示．

再选择一个子问题进行探测，如选(CP_2)．放弃(CP_2)中的整数要求，解相应的线性规划，得知无可行解．于是由算法知在这节点无须再分支，其树形图如图2-3所示．

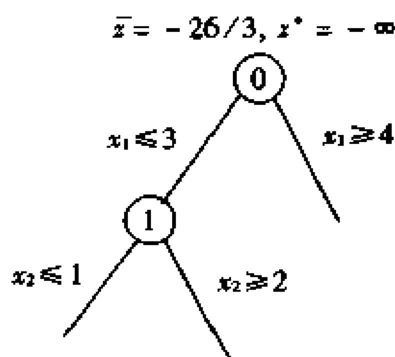


图 2-2

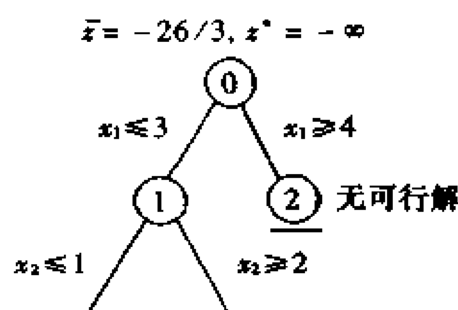


图 2-3

继续选择有待探测的一个子问题(节点)，如在(P)基础上加上约束：$x_1 \leqslant 3$ 和 $x_2 \leqslant 1$ 的子问题．松弛后求得相应线性规划无可行解．因此无须再分支．相应树形图如图2-4所示．

从图2-4上看出，在 Π 中尚有一个有待探测的子问题，再求其相应松弛后的线性规划，得最优解 $x_1 = 2, x_2 = 2$，最优值 $z = -10$．由于最优解为整数解，因此根据算法，记录此原问题(P)的可行解并修改 z^*，$z^* = -10$，无须再分支．树形图改为图2-5．

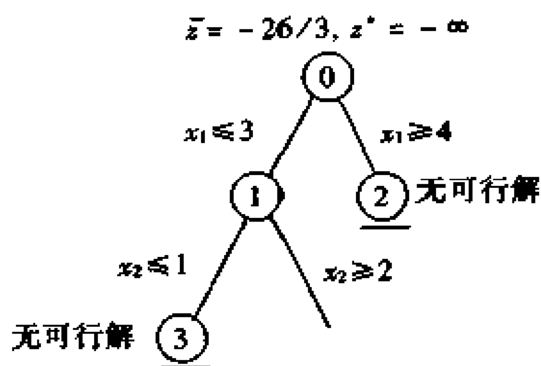


图 2-4

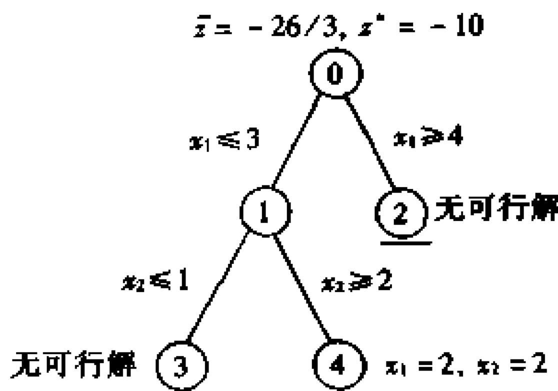


图 2-5

由图 2-5 可知,已无有待探测的子问题,即算法步骤中的 $\Pi=\varnothing$,因而算法终止,已求得(P)的最优值为 $z^*=-10$,相应的可行解 $x_1^*=2, x_2^*=2$,即为(P)的最优解.

2.3 割平面法

线性整数规划(P)放弃整数性要求后成为线性规划($\tilde{\mathrm{P}}$).前面在理论上讨论了($\tilde{\mathrm{P}}$)可行域的极点都是整数点的条件下的求解方法.如果不具备这个条件,($\tilde{\mathrm{P}}$)的最优解就不一定具备整数性条件,这时需进一步处理即改进($\tilde{\mathrm{P}}$).割平面法便是改进($\tilde{\mathrm{P}}$)的一种方法,即再加上适当的线性约束 —— 割约束,逐步把(P)的最优解变为($\tilde{\mathrm{P}}$)可行域的极点和除去所有非整的、具有更好目标函数值的极点.

2.3.1 基本割平面

设线性整数规划为

(P)
$$\max \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}=x_0, \tag{2-1}$$
$$\text{s.t.}\quad A\boldsymbol{x}=\boldsymbol{b}, \boldsymbol{x}\geqslant \boldsymbol{0}, \text{且为整数},$$

并设其中数据均为整数,A 为 $m\times n$ 矩阵,$r(A)=m$.

($\tilde{\mathrm{P}}$)
$$\max \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \tag{2-2}$$
$$\text{s.t.}\quad A\boldsymbol{x}=\boldsymbol{b}, \boldsymbol{x}\geqslant \boldsymbol{0}.$$

假设已把 $A\boldsymbol{x}=\boldsymbol{b}$ 解成

$$x_{B_i}=y_{i0}-\sum_{j\in R}y_{ij}x_j,\quad i=0,1,2,\cdots,m, \tag{2-3}$$

则由(2-3)式给出的基本解为

$$x_{B_i}=y_{i0},\quad i=0,1,2,\cdots,m,\quad x_j=0, j\in R,$$

其中 R 为非基本变量集.

用 $h\neq 0$ 乘(2-3)式,得

$$hx_{B_i}+\sum_{j\in R}hy_{ij}x_j=hy_{i0},$$

因要求 $\boldsymbol{x}\geqslant 0$,所以

$$[h]x_{B_i}+\sum_{j\in R}[hy_{ij}]x_j\leqslant hy_{i0}. \tag{2-4}$$

又因要求 $\boldsymbol{x}$ 为整数,所以(2-4)式左边为整数,因此有

$$[h]x_{B_i}+\sum_{j\in R}[hy_{ij}]x_j\leqslant [hy_{i0}], \tag{2-5}$$

用[h]乘(2-3)式再减(2-5)式,得

$$\sum_{j\in R}([h]y_{ij}-[hy_{ij}])x_j\geqslant [h]y_{i0}-[hy_{i0}]. \tag{2-6}$$

若由(2-3)式给出的基本解不是整的就不能满足(2-6)式.而任何(P)的可行解均能满足(2-6)式,(2-6)式称为**基本割平面**.用不同的方法应用(2-6)式就得到各种不同的割平面及割平面算法.

2.3.2　分数割平面法

在基本割平面(2-6)式中,令 $h = 1$,则可得

$$\sum_{j\in R}(y_{ij} - [y_{ij}])x_j \geqslant y_{i0} - [y_{i0}],$$

记 $r_{ij} = y_{ij} - [y_{ij}], j = 0$ 和 $j \in R$,则上式可写为

$$\sum_{j\in R} r_{ij}x_j \geqslant r_{i0},$$

引进松弛变量 S_i,得条件

$$S_i = - r_{i0} + \sum_{j\in R} r_{ij}x_j, \tag{2-7}$$

称(2-7)式为**分数割平面**.因为

$$\begin{aligned} x_i &= ([y_{i0}] - \sum_{j\in R}[y_{ij}]x_j) - (- r_{i0} + \sum_{j\in R} r_{ij}x_j) \\ &= ([y_{i0}] - \sum_{j\in R}[y_{ij}]x_j) - S_i, \end{aligned}$$

所以,当 x_i, x_j 都取非负整数时,S_i 自然也取非负整数值.

分数割平面法求解线性整数规划(P)的步骤如下:

(1) 用对偶单纯形算法求解线性规划松弛问题($\tilde{\mathrm{P}}$).若($\tilde{\mathrm{P}}$)无最优解,则(P)也无最优解,算法终止;若($\tilde{\mathrm{P}}$)有最优解,则设

$$\bar{x} = y_0 + \sum_{j\in R} y_j(- x_j),$$

其中 $y_{i0} \geqslant 0, y_{0j} \geqslant 0, i = 1,2,\cdots,n, j \in R$.

(2) 若所有 y_{i0} 都是整数,则算法终止,$x^* = y_0$ 即为(P)的最优解;相反,设 y_{l0} 是非整数,取诱导方程为

$$x_l = y_{l0} + \sum_{j\in R} y_{lj}(- x_j)$$

(即取诱导行为 l 行).

(3) 从诱导方程导出分数割平面

$$S_l = - r_{l0} - \sum_{j\in R} r_{lj}(- x_j),$$

其中 $r_{lj} = y_{lj} - [y_{lj}]$.

(4) 计算

$$\max\left\{\frac{- y_{0j}}{- r_{lj}}\,\middle|\, r_{lj} > 0, j \in R\right\} = - \frac{y_{0S}}{r_{lS}}.$$

(5) 用 S_l 替换非基变量 x_S,再求改进后的松弛问题的最优解.返回(2).

事实上引进割平面条件,只是作一次参数变换,并不需要把割平面条件真的加

入约束方程组中去.

例2　求解整数规则：

$$\min(4x+5y),$$
$$\text{s.t.}\ 3x+2y\geqslant 7,\ x+4y\geqslant 5,$$
$$3x+y\geqslant 2,$$
$$x,y\geqslant 0\ 且为整数.$$

写成参数形式为

$$\max x_0,$$
$$\text{s.t.}\ x_0=-4x_4-5x_5,\ x_1=-7+3x_4+2x_5,$$
$$x_2=-5+x_4+4x_5,\ x_3=-2+3x_4+x_5,$$
$$x_1,x_2,x_3,x_4,x_5\geqslant 0,$$
$$x_i\ 均为整数,\ i=0,1,\cdots,5.$$

解　用对偶单纯形法解松弛后的线性规划，得

	=	$-x_1$	$-x_2$
x_0	$-112/10$	$11/10$	$7/10$
x_1	0	-1	0
x_2	0	0	-1
x_3	$42/10$	$-11/10$	$3/10$
x_4	$18/10$	$-4/10$	$2/10$
x_5	$8/10$	$1/10$	$-3/10$

因 $y_{00}=-112/10$ 不是整数，选诱导方程为

$$x_0=-\frac{112}{10}-\frac{11}{10}x_1-\frac{7}{10}x_2,$$

导出割平面

$$S_0=-\frac{8}{10}-\frac{1}{10}(-x_1)-\frac{7}{10}(-x_2),$$

按步骤(4)求

$$\max\left\{\frac{\frac{11}{10}}{\left(-\frac{1}{10}\right)},\frac{\frac{7}{10}}{\left(-\frac{7}{10}\right)}\right\}=-1,$$

得 $S=2$. 以 S_0 替换 x_2，得表达式为

	=	$-x_1$	$-S_0$
x_0	-12	1	1
x_1	0	-1	0
x_2	$8/7$	$1/7$	$-10/7$
x_3	$27/7$	$-8/7$	$3/7$
x_4	$11/7$	$-3/7$	$2/7$
x_5	$8/7$	$1/7$	$-3/7$

这已是线性规划最优解，但 $y_{20}=8/7$ 不是整数，选诱导方程为

$$x_2=\frac{8}{7}+\frac{1}{7}(-x_1)-\frac{10}{7}(-S_0),$$

导出割平面

$$S_2=-\frac{1}{7}-\frac{1}{7}(-x_1)-\frac{4}{7}(-S_0),$$

求

$$\max\left\{\frac{1}{-\left(\frac{1}{7}\right)},\frac{1}{\left(-\frac{4}{7}\right)}\right\}=-\frac{7}{4},$$

得 $S=2$. 以 S_2 替换 S_0，得表示式为

	=	$-x_1$	$-S_2$
x_0	$-121/4$	3/4	7/4
x_1	0	-1	0
x_2	6/4	2/4	$-10/4$
x_3	15/4	$-5/4$	3/4
x_4	6/4	$-2/4$	2/4
x_5	5/4	1/4	$-3/4$

又得出最优解，但 $y_{00}=-121/4$ 不是整数，选诱导方程为

$$x_0=-\frac{121}{4}+\frac{3}{4}(-x_1)+\frac{7}{4}(-S_2),$$

导出割平面为

$$S'_0=-\frac{3}{4}-\frac{3}{4}(-x_1)-\frac{3}{4}(-S_2),$$

求

$$\max\left\{\frac{\frac{3}{4}}{\left(-\frac{3}{4}\right)},\frac{\frac{7}{4}}{\left(-\frac{3}{4}\right)}\right\}=-1,$$

得 $S=1$. 以 S'_0 替换 x_1，得表示式为

	=	$-S'_0$	$-S_2$
x_0	-13	1	1
x_1	1	$-4/3$	1
x_2	1	2/3	-3
x_3	5	$-5/3$	2
x_4	2	$-2/3$	1
x_5	1	1/3	-1

已得线性规划最优解，且又都是整数，故 $x^*=x_4=2, y^*=x_5=1$ 即为原线性整数规划的最优解，最优值为 $x_0^*=13$.

此例计算过程如图 2-6 中点 ①→②→③→④→⑤→⑥.

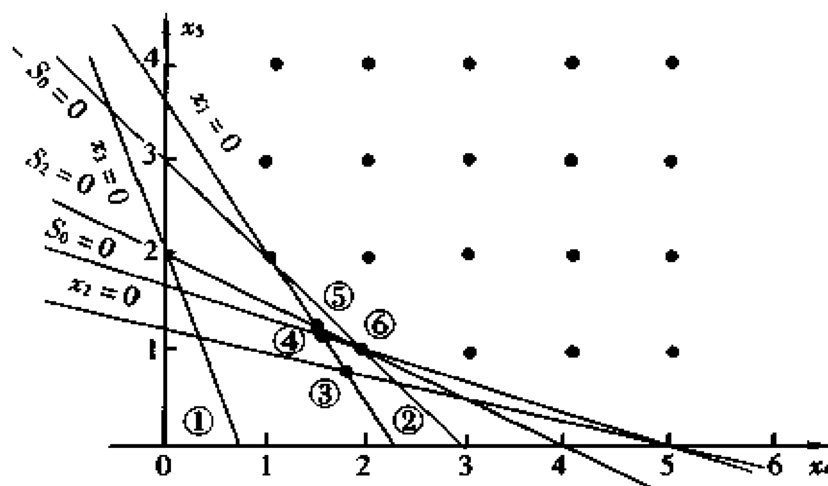


图 2-6

第一个割平面为

$$S_0 = -\frac{8}{10} + \frac{1}{10}x_1 + \frac{7}{10}x_2 = -5 + x_4 + 3x_5 \geqslant 0,$$

加上它后,使点 ③ → 点 ④.

第二个割平面为

$$S_2 = -\frac{1}{7} + \frac{1}{7}x_1 + \frac{4}{7}S_0 = -4 + x_4 + 2x_5 \geqslant 0,$$

加上它后,使点 ④ → 点 ⑤.

第三个割平面为

$$S'_0 = -\frac{3}{4} + \frac{3}{4}x_1 + \frac{3}{4}S_2 = -9 + 3x_4 + 3x_5 \geqslant 0,$$

加上它后,使点 ⑤ → 点 ⑥.

2.3.3　对偶整数割平面算法

在基本割平面((2-6)式)中,令 $h = \frac{1}{\lambda}$(λ 为正整数),可导出割平面条件

$$\left[\frac{y_{i0}}{\lambda}\right] + \sum_{j\in R}\left[\frac{y_{ij}}{\lambda}\right](-x_j) \geqslant \left[\frac{1}{\lambda}\right]y_{i0} + \sum_{j\in R}\left[\frac{1}{\lambda}\right]y_{ij}(-x_j) = \left[\frac{1}{\lambda}\right]x_i \geqslant 0,$$

引进松弛变量 S_i 后,可得

$$S_i = \left[\frac{y_{i0}}{\lambda}\right] + \sum_{j\in R}\left[\frac{y_{ij}}{\lambda}\right](-x_j) \geqslant 0, \tag{2-8}$$

显然,当 x_j 是整数时,S_i 也是整数.称(2-8)式为对偶整数割平面.

对偶整数割平面法步骤如下:

假设求($\tilde{\mathrm{P}}$)时已得到表达式

$$\bar{x} = y_0 + \sum_{j\in R}y_j(-x_j),$$

且满足

1) y_{ij} 都是整数;

2) $y_{0j} \geqslant 0 (j \in R)$.

(1) 若所有 $y_{i0} \geqslant 0$, 则算法终止, $\bar{x} = y_0$ 即为(P)的最优解;相反,设有某 $y_{l0} < 0 (1 \leqslant l \leqslant n)$,取诱导方程

$$x_l = y_{l0} + \sum_{j \in R} y_{lj}(-x_j).$$

(2) 若所有 $y_{lj} \geqslant 0 (j \in R)$,则算法终止,(P) 无可行解;相反,求

$$\min\{y_{0j} \mid y_{lj} < 0, j \in R\} = y_{0S}.$$

(3) 取 $\lambda = \max\{|y_{lj}| \mid y_{lj} < 0, j \in R\}$.

(4) 导出整数割平面

$$S_l = \left[\frac{y_{l0}}{\lambda}\right] + \sum_{j \in R} \left[\frac{y_{lj}}{\lambda}\right](-x_j).$$

(5) 以 S_l 替换 x_S,得表达式

$$\bar{x} = \bar{y}_0 + \sum_{j \in R} \bar{y}_j(-\bar{x}_j),$$

其中 $\bar{y}_S = \bar{y}_S, \bar{y}_j = \bar{y}_j + \left[\frac{y_{lj}}{\lambda}\right] y_S (j \neq S), \bar{x}_S = S_l, \bar{x}_j = x_j (j \neq S)$. 返回(1).

下面用对偶整数割平面法求解例2.对此例只要加上松弛变量后,就可得到满足1)和2)的表达式

	=	$-x_4$	$-x_5$
x_0	6	4	5
x_1	-7	-3	-2
x_2	-5	-1	-4
x_3	-2	-3	-1
x_4	0	-1	0
x_5	0	0	-1

因 $y_{10} = -7 < 0$,诱导方程为

求

$$x_1 = -7 - 3(-x_4) - 2(-x_5).$$

取

$$\min\{4, 5\} = 4 = y_{01},$$

$$\lambda = \max\{|-3|, |-2|\} = 3,$$

导出割平面

$$S_1 = \left[\frac{-7}{3}\right] + \left[\frac{-3}{3}\right](-x_4) + \left[\frac{-2}{3}\right](-x_5)$$

$$= -3 - (-x_4) - (-x_5).$$

以 S_1 替换 x_4,得表达式为

	=	$-S_1$	$-x_5$
x_0	−12	4	1
x_1	2	−3	1
x_2	−2	−1	−3
x_3	7	−3	2
x_4	3	−1	1
x_5	0	0	−1

因 $y_{20}=-2<0$,诱导方程为

求 $$x_2=-2-(-S_1)-3(-x_5).$$

取 $$\min\{4,1\}=1=y_{02},$$

$$\lambda=\max\{|-1|,|-3|\}=3,$$

导出割平面

$$S_2=\left[\frac{-2}{3}\right]+\left[\frac{-1}{3}\right](-S_1)+\left[\frac{-3}{3}\right](-x_5)$$
$$=-1-(-S_1)-(-x_5).$$

以 S_2 替换 x_5,得表达式为

	=	$-S_1$	$-S_2$
x_0	−13	3	1
x_1	1	−4	1
x_2	1	2	−3
x_3	5	−5	2
x_4	2	−2	1
x_5	1	1	−1

因所有 $y_{i0}\geqslant 0(1\leqslant i\leqslant 5)$,故已得例 2 的最优解为 $x_1^*=x_2^*=1,x_3^*=5,x_4^*=2,x_5^*=1$,最优值为 $x_0^*=13$.

计算过程如图 2-7 中点 ① → 点 B → 点 C.

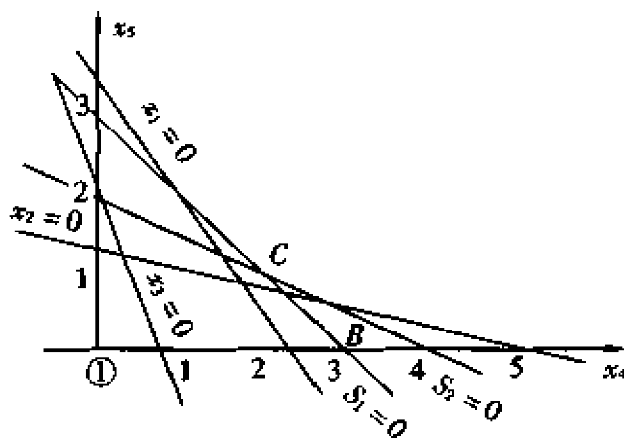


图 2-7

2.4　0-1 规划的隐枚举法

例 3　已知 0-1 规划模型为

$$\max z = 3x_1 - 2x_2 + 5x_3,$$

$$\text{s.t.}\quad x_1 + 2x_2 - x_3 \leqslant 2, \tag{2-9}$$

$$x_1 + 4x_2 + x_3 \leqslant 4, \tag{2-10}$$

$$x_1 + x_2 \leqslant 3, \tag{2-11}$$

$$4x_2 + x_3 \leqslant 6, \tag{2-12}$$

$$x_1, x_2, x_3 = 0 \text{ 或 } 1.$$

求解思路及改进措施如下：

(1) 先试探性求一个可行解. 易看出 $(x_1, x_2, x_3) = (1,0,0)$ 满足 (2-9) ~ (2-12) 式的约束条件，故为一个可行解，且 $z = 3$.

(2) 因为是求最大值，故求最优解时，凡是目标值 $z < 3$ 的解不必检验是否满足约束条件即可删掉，因为它肯定不是最优解，于是可增加一个约束条件（目标值下界）

$$3x_1 - 2x_2 + 5x_3 \geqslant 3. \tag{2-13}$$

该式称为过滤条件，凡是与检验点相应的目标函数值 $z < 3$，全被滤掉，故称隐枚举法.

本例采用隐枚举法的求解过程示于表 2-1 中. 从表中看出，因有 3 个自变量，故有 $2^3 = 8$ 种组合，每种组合需计算目标值和 4 个约束，按穷举法需计算 $5 \times 8 = 40$ 次，才能得出最优解. 而按隐枚举法，对每种组合逐一检查约束，当出现违反约束时立即停止，转向计算下一个，对本例只计算 16 次即可. 其最优解为 $(x_1^*, x_2^*, x_3^*) = (1,0,1)$，$z^* = 8$.

表 2-1

组合点 (x_1, x_2, x_3)	条件					是否满足约束	
	z	(1)	(2)	(3)	(4)		
(0,0,0)	0					×	
(0,0,1)	5	−1	1	0	1	√	此时可立即将过滤条件 (2-13) 改为 > 5.
(0,1,0)	−2					×	
(0,1,1)	3					×	
(1,0,0)	3					×	
(1,0,1)	8	0	2	1	1	√	又可将 (2-13) 式改为 > 8.
(1,1,0)	1					×	
(1,1,1)	6					×	

(3) 由于对每个组合首先计算目标值以验证过滤条件，故应优先计算目标值 z

大的组合,这样可提前抬高过滤门槛,以减少计算量.于是组合变量 x_1,x_2,x_3 在采用表 2-1 形式时,其排列顺序应按目标函数中系数递增顺序排列,如在例 3 中由 z 的表达式可知x_2的系数以 -2 为最小,x_3 的系数以 5 为最大,故排为(x_2,x_1,x_3).改进后的计算见表 2-2.

表 2-2

组合点(x_2,x_1,x_3)	条件					是否满足约束	
	z	(1)	(2)	(3)	(4)		
(0,0,0)	0					×	
(0,0,1)	5	-1	1	0	1	√	(2-13)式改为 >5
(0,1,0)	3					×	
(0,1,1)	8	0	2	1	1	√	(2-13)式改为 >8
(1,0,0)	-2					×	
(1,0,1)	3					×	
(1,1,0)	1					×	
(1,1,1)	6					×	

3　线性混合整数规划解法

对于线性混合整数规划,分支定界法和割平面法都是适用的.例如,分支定界法只需对要求取整的变量进行分解和实施其他的原则.关于割平面法也是这样的原则,其步骤如下:

对于线性混合整数规划(P),设它的松弛线性规划为$(\tilde{P})$,且$(\tilde{P})$解的表达式为

$$\bar{x} = \bar{y}_0 + \sum_{j\in R} y_j(-x_j).$$

假设在(P)中,要求取整数值的变量为 $x_1,x_2,\cdots,x_r$.因此,若 $y_{10},y_{20},\cdots,y_{r0}$ 都是整数,则 $\bar{x}=y_0$ 即为(P)的最优解;相反,设有 $l(1\leqslant l\leqslant r)$,使 y_{l0} 不是整数,取诱导方程

$$x_l = y_{l0} + \sum_{j\in R} y_{lj}(-x_j),$$

按是否为整数变量,将 $x_j(j\in R)$ 分为两点,即

$$I = \{j \mid x_j \text{为整数变量}, j\in R\}.$$

设

$$J = \{j \mid x_j \text{不是整数变量}, j\in R\},$$

$$y_{l0} = [y_{l0}] + r_{l0},$$

$$y_{lj} = [y_{lj}] + r_{lj}, j\in I,$$

以及

$$I_1 = \{j \mid j \in I, r_{lj} \leqslant r_{l0}\},$$
$$I_2 = \{j \mid j \in I, r_{lj} > r_{l0}\},$$
$$J_1 = \{j \mid j \in J, y_{lj} \geqslant 0\},$$
$$J_2 = \{j \mid j \in J, y_{lj} < 0\},$$
$$f_{l0} = r_{l0},$$
$$f_{lj} = r_{lj}, j \in I_1,$$
$$f_{lj} = \frac{r_{lj} - 1}{r_{l0} - 1} r_{l0}, j \in I_2,$$
$$f_{lj} = y_{lj}, j \in J_1,$$
$$f_{lj} = \frac{y_{lj}}{r_{l0-1}} r_{l0}, j \in J_2,$$

则(P)的任何可行解,必使

$$S_l^* = -f_{l0} + \sum_{j \in R} (-f_{lj})(-x_j) \geqslant 0, \tag{3-1}$$

称(3-1)式为线性混合整数规划的割平面.

相应割平面算法的步骤如下:

(1) 解($\tilde{\mathrm{P}}$) 若($\tilde{\mathrm{P}}$)没有最优解,则终止,(P)也没有最优解;若($\tilde{\mathrm{P}}$)有最优解,其表达式为

$$\bar{x} = y_0 + \sum_{j \in R} y_j (-x_j).$$

(2) 若 y_{i0} 都是整数,$i = 1,2,\cdots,r$,则终止,$\bar{x} = y_0$ 即为(P)的最优解;相反,设有某 $y_{l0}(l \leqslant r)$ 不是整数,取诱导方程

$$x_l = y_{l0} + \sum_{j \in R} y_{lj} (-x_j).$$

(3) 导出混合割平面

$$S_l^* = -f_{l0} - \sum_{j \in R} f_{lj} (-x_j).$$

(4) 用 S_l^* 替换 x_S,得新的表达式

$$\bar{\bar{x}} = \bar{y}_0 + \sum_{j \in R} \bar{y}_j (-\bar{x}_j),$$

其中,$\bar{x}_j = x_j (j \neq S)$,$\bar{x}_S = S_l^*$,$\bar{y}_S = \frac{1}{f_{lS}} y_S$,$\bar{Y}_j = y_j - \frac{f_{lj}}{f_{lS}} y_S (j \neq S)$.

(5) 若 $\bar{y}_{i0} \geqslant 0 (i = 1,2,\cdots,n)$,返回(2);否则,返回(1).

例 1 求

$$\max x_0 = -4x_2 - 10x_4 + 20,$$
$$\text{s.t.} \quad x_1 - \frac{5}{3}x_2 - \frac{1}{3}x_4 = \frac{5}{3},$$
$$-\frac{4}{3}x_2 + x_3 + \frac{11}{3}x_4 = 7/3,$$

$x_1, x_2, x_3, x_4 \geqslant 0$, x_3, x_4 取整数值.

松弛问题的最优解为

		$-x_2$	$-x_4$
x_0	20	4	10
x_3	7/3	-4/3	11/3
x_4	0	0	-1
x_1	5/3	-5/3	-1/3
x_2	0	-1	0

取诱导方程

$$x_3 = \frac{7}{3} - \frac{4}{3}(-x_2) + \frac{11}{3}(-x_4),$$

$$I = \{4\}, J = \{2\}, r_{30} = \frac{1}{3}, r_{34} = \frac{2}{3}.$$

$$I_1 = \varnothing, I_2 = \{4\}, J_1 = \varnothing, J_2 = \{2\}.$$

因此 $f_{30} = \frac{1}{3}, f_{32} = \frac{-4/3}{1/3-1} \times \frac{1}{3} = \frac{2}{3}, f_{34} = \frac{2/3-1}{1/3-1} \times \frac{1}{3} = \frac{1}{6}$.

混合割平面为

$$S_3^* = -\frac{1}{3} - \frac{2}{3}(-x_2) - \frac{1}{6}(-x_4),$$

$$\max\left\{\frac{4}{-\frac{2}{3}}, \frac{10}{-\frac{1}{6}}\right\} = -6,$$

得 $S = 2$.用 S_3^* 代替 x_2 后,得表达式

		$-S_3^*$	$-x_4$
x_0	18	6	9
x_3	3	-2	4
x_4	0	0	-1
x_1	15/6	-5/2	1/12
x_2	1/2	-3/2	1/4

由于 $y_{i0} \geqslant 0$,且 y_{30}, y_{40} 均为整数,所以(P)的最优解为 $x_1^* = 15/6, x_2^* = 1/2, x_3^* = 3, x_4^* = 0$,最优值为 $x_0^* = 18$.

3.1 拉格朗日松驰法

假设线性混合整数规划,已被分离成如下形式:

(P) $$\max x_0 = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y},$$

$$\text{s.t. } \boldsymbol{A}_{11}\boldsymbol{x} + \boldsymbol{A}_{12}\boldsymbol{y} = \boldsymbol{b}_1,$$

$$\boldsymbol{A}_{21}\boldsymbol{x} + \boldsymbol{A}_{22}\boldsymbol{y} = \boldsymbol{b}_2,$$

$$\boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}, \boldsymbol{y} \text{ 为整数向量}.$$

其中 $A_{11}, A_{12}, A_{21}, A_{22}$ 都是具有一定维数的矩阵,对于任意给定的行向量 $\boldsymbol{\Pi}_2$(其维

数与 b_2 的维数相同),定义(P)的松弛问题如下:

$$\mathrm{R}(\boldsymbol{\Pi}_2)\qquad \max x_0(\boldsymbol{\Pi}_2) = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{\Pi}_2(\boldsymbol{b}_2 - A_{21}\boldsymbol{x} - A_{22}\boldsymbol{y})$$
$$= (\boldsymbol{c}^{\mathrm{T}} + \boldsymbol{\Pi}_2 A_{21})\boldsymbol{x} + (\boldsymbol{d}^{\mathrm{T}} + \boldsymbol{\Pi}_2 A_{22})\boldsymbol{y} - \boldsymbol{\Pi}_2\boldsymbol{b}_2,$$
$$\text{s.t.}\quad A_{11}\boldsymbol{x} + A_{12}\boldsymbol{y} = \boldsymbol{b}_1,$$
$$\boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}, \boldsymbol{y}\ \text{为整数向量}.$$

称 $\mathrm{R}(\boldsymbol{\Pi}_2)$ 为(P)的一个**拉格朗日(Lagrange)松弛问题**.假如松弛问题 $\mathrm{R}(\boldsymbol{\Pi}_2)$ 的最优解 $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ 还满足

$$A_{21}\boldsymbol{x}^* + A_{22}\boldsymbol{y}^* = \boldsymbol{b}_2,$$

那么,它必是(P)的最优解.

拉格朗日松弛法是一种比较实用的整数规划的近似算法,其前提是假设 $\mathrm{R}(\boldsymbol{\Pi}_2)$ 比(P)容易求解,且都有最优解.其步骤为:

(1) 任给一初始的 $\boldsymbol{\Pi}_2$(例如零向量).

(2) 求解 $\mathrm{R}(\boldsymbol{\Pi}_2)$,设最优解为 $(\boldsymbol{x}^*, \boldsymbol{y}^*)$,最优值为 $x_0^*(\boldsymbol{\Pi}_2)$.

(3) 若 $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ 满足

$$A_{21}\boldsymbol{x}^* + A_{22}\boldsymbol{y}^* = \boldsymbol{b}_2$$

或

$$\| A_{21}\boldsymbol{x}^* + A_{22}\boldsymbol{y}^* - \boldsymbol{b}_2 \| \leqslant \epsilon,$$

则步骤终止,$(\boldsymbol{x}^*, \boldsymbol{y}^*)$ 即为(P)的最优解或近似最优解;相反,转(4).

(4) 以 $\boldsymbol{\Pi}_2 + t^0(\boldsymbol{b}_2 - A_{21}\boldsymbol{x}^* - A_{22}\boldsymbol{y}^*)^{\mathrm{T}}$ 代替原来的 $\boldsymbol{\Pi}_2$,返回(2).其中 t^0 是一个可变的步长,通常取为

$$t^0 = \frac{\lambda(x_0^*(\boldsymbol{\Pi}_2) - z^*)}{\| \boldsymbol{b}_2 - A_{21}\boldsymbol{x}^* - A_{22}\boldsymbol{y}^* \|},$$

其中 z^* 是问题(P)的最优值的一个下界估计值,λ 是可变的比例数,开始时,可取 $\lambda = 2$,在迭代过程中,感到 $x_0^*(\boldsymbol{\Pi}_2)$ 下降太慢时,可逐次将 λ 缩小 $\frac{1}{2}$ 倍.

例 2　选址问题

设有 n 个城市 $\{1,2,\cdots,n\}$,每日需要某种物资,计划要在其中的 m 个城市中建造 m 座生产这种物资的工厂,假设各工厂的规模不限.设 c_{ij} 表示若在城市 i 建厂,而城市 j 完全由城市 i 负责供应时的总运费.定义 0-1 变量 x_i 及 x_{ij} 如下:

$$x_i = \begin{cases} 1, & \text{若在城市 } i \text{ 建厂;} \\ 0, & \text{若不在城市 } i \text{ 建厂.} \end{cases}$$

$$x_{ij} = \begin{cases} 1, & \text{城市 } j \text{ 由城市 } i \text{ 供应;} \\ 0, & \text{城市 } i \text{ 由城市 } j \text{ 供应,} \end{cases}$$

则问题可写成如下形式:

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij}\ (\text{总运费最省}),$$

$$\text{s.t.}\ \sum_{i=1}^{n} x_{ij} = 1, j = 1,2,\cdots,n,$$

$$\sum_{i=1}^{n} x_i = m,$$
$$x_{ij} \leqslant x_i, i, j = 1,2,\cdots,n,$$
$$x_i, x_{ij} = 0,1.$$

取拉格朗日松弛问题为求

$$\min\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} \Pi_j x_{ij} - \sum_{j=1}^{n} \Pi_j\right\},$$
$$\text{s.t.} \sum_{i=1}^{n} x_i = m,$$
$$x_{ij} \leqslant x_i, i, j = 1,2,\cdots,n,$$
$$x_i, x_{ij} = 0,1.$$

易见,松弛问题的最优解必须满足关系

(II):

$$x_{ij} = \begin{cases} x_i, & \text{当 } c_{ij} + \Pi_j < 0 \text{ 时;} \\ 0, & \text{当 } c_{ij} + \Pi_j > 0 \text{ 时;} \\ x_i \text{ 或 } 0, & \text{当 } c_{ij} + \Pi_j = 0 \text{ 时.} \end{cases}$$

设

$$\overline{c_{ij}} = \min\{c_{ij} + \Pi_j, 0\},$$
$$\overline{c_i} = \sum_{j=1}^{n} \overline{c_{ij}},$$

则松弛问题可以等价地写成如下简单形式:

$$\min \sum_{i=1}^{n} \overline{c_i} x_i,$$
$$\text{s.t.} \sum_{i=1}^{n} x_i = m, x_i = 0,1.$$

不妨设$\overline{c_1} \leqslant \overline{c_2} \leqslant \cdots \leqslant \overline{c_n}$,则上述问题的最优解为

$$x_1 = x_2 = \cdots = x_m = 1, \text{其余 } x_j = 0,$$

再由关系(II),确定松弛问题的最优解(x_i^*, x_{ij}^*).

假如,这时的(x_i^*, x_{ij}^*)已满足

$$\sum_{i=1}^{n} x_{ij}^* = 1, \quad j = 1,2,\cdots,n,$$

则已求得选址问题的最优解.相反,用

$$\Pi_j + t_0\left(\sum_{i=1}^{n} x_{ij}^* - 1\right)$$

代替原来的 Π_j,从而得到新的拉格朗日松弛问题,其中

$$t_0 = \frac{\lambda \left| \sum_{j=1}^{n} \bar{c}_j x_j^* - z^* \right|}{\sum_{j=1}^{n} \left(\sum_{i=1}^{n} x_{ij}^* - 1 \right)^2},$$

$0 \leqslant \lambda \leqslant 2$, z^* 为某个选址方案的目标函数值.

3.2 交叉分解算法

交叉分解算法是将分解算法和拉格朗日松弛法相结合的一种方法,它也是一种比较实用的近似算法.

(P):关于混合整数规划

$$\begin{aligned} \max x_0 &= \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y}, \\ \text{s.t. } & \boldsymbol{A}_{11}\boldsymbol{x} + \boldsymbol{A}_{12}\boldsymbol{y} = \boldsymbol{b}_1, \\ & \boldsymbol{A}_{21}\boldsymbol{x} + \boldsymbol{A}_{22}\boldsymbol{y} = \boldsymbol{b}_2. \\ & \boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}, \boldsymbol{y} \text{ 为整数向量}. \end{aligned}$$

从分解算法中,对应于给定的整数向量 $\boldsymbol{y}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}$,定义子规划

$$\begin{aligned} x_0(\boldsymbol{y}) &= \max\left\{ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} \,\middle|\, \begin{pmatrix} \boldsymbol{A}_{11} \\ \boldsymbol{A}_{21} \end{pmatrix} \boldsymbol{x} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{A}_{12} \\ \boldsymbol{A}_{22} \end{pmatrix} \boldsymbol{y}, \boldsymbol{x} \geqslant \boldsymbol{0} \right\} \\ &= \min\{ \boldsymbol{\Pi}_1\boldsymbol{b}_1 + \boldsymbol{\Pi}_2\boldsymbol{b}_2 - (\boldsymbol{\Pi}_1\boldsymbol{A}_{12} + \boldsymbol{\Pi}_2\boldsymbol{A}_{22})\boldsymbol{y} + \\ & \qquad \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} \mid \boldsymbol{\Pi}_1\boldsymbol{A}_{11} + \boldsymbol{\Pi}_2\boldsymbol{A}_{21} \geqslant \boldsymbol{c}^{\mathrm{T}} \}. \end{aligned}$$

由拉格朗日松弛法,对应于乘子 $\boldsymbol{\Pi}_2$,可以定义拉格朗日松弛问题 $\mathrm{R}(\boldsymbol{\Pi}_2)$:

$$\begin{aligned} x_0(\boldsymbol{\Pi}_2) = \max\{ & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{y} - \Pi_2(\boldsymbol{b}_2 - \boldsymbol{A}_{21}\boldsymbol{x} - \boldsymbol{A}_{22}\boldsymbol{y}) \mid \boldsymbol{A}_{11}\boldsymbol{x} + \\ & \boldsymbol{A}_{12}\boldsymbol{y} = \boldsymbol{b}_1, \boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}, \boldsymbol{y} \text{ 为整数向量} \}. \end{aligned}$$

对上述的任何子规划和松弛问题,必有关系

$$x_0(\boldsymbol{\Pi}_2) \geqslant x_0(\boldsymbol{y}),$$

当等式成立时,子规划的最优解便是原问题(P)的最优解.

交叉算法的基本思想,是在分解算法中,用求解拉格朗日松弛问题来代替求解松弛问题,为子问题修正参数 $\boldsymbol{y}$,或者说,在拉格朗日松弛算法中,用求解子规划来修正乘子 $\boldsymbol{\Pi}_2$.

下面假设**对偶子规划**的可行域

$$\{ \boldsymbol{\Pi}_1\boldsymbol{A}_{11} + \boldsymbol{\Pi}_2\boldsymbol{A}_{21} \geqslant \boldsymbol{c}^{\mathrm{T}} \}$$

和拉格朗日松弛问题的可行域

$$\{ \boldsymbol{A}_{11}\boldsymbol{x} + \boldsymbol{A}_{12}\boldsymbol{y} = \boldsymbol{b}_1, \boldsymbol{x} \geqslant \boldsymbol{0}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}, \boldsymbol{y} \text{ 为整数向量} \},$$

都非空有界.

交叉分解算法的步骤为:

(1) 任取一整数向量 $\boldsymbol{y}, \boldsymbol{e} \geqslant \boldsymbol{y} \geqslant \boldsymbol{0}$,用单纯形法解子规划 $x_0(\boldsymbol{y})$. 设求得(对偶)的最优解为 $\boldsymbol{\Pi}^1 = (\Pi_1^1, \Pi_2^1)$. 置 $Q = \{\boldsymbol{\Pi}^1\}$, $R = \varnothing$, $i = 1$.

(2) 求解拉格朗日松弛问题 $\mathrm{R}(\boldsymbol{\Pi}_2^i)$,记其最优解为$(\boldsymbol{x}^*, \boldsymbol{y}^*)$.

(3) 求解子规划 $x_0(y^*)$,设其(对偶)最优解为 $\Pi^* = (\Pi_1^*, \Pi_2^*)$.

(4) 若 $\Pi_2^* = \Pi_2^i$,则终止,子规划 $x_0(y^*)$ 的最优解也是原问题(P)的最优解;相反,转(5).

(5) 下述情形任选其一:

1° 置 $\Pi^{i+1} = \Pi^*$, $Q \cup \{\Pi^{i+1}\} \to Q$, $i+1 \to i$,返回(2);

2° 转(6).

(6) 求解松弛主规划

P(Q,R)

$$x_0(Q,R) = \max x_0,$$

$$\text{s.t.}\ d^{\mathrm T}y + \Pi_1^i b_1 + \Pi_2^i b_2 - (\Pi_1^i A_{12} + \Pi_2^i A_{22})y \geqslant x_0,$$

(对所有的 $\Pi^i \in Q$)

$$0 \leqslant y \leqslant e,\quad y\ \text{是整数向量}.$$

设其最优解为 y^*.

(7) 解子规划 $x_0(y^*)$,设其(对偶)最优解为 $\Pi^* = (\Pi_1^*, \Pi_2^*)$.

(8) 若 $x_0(y^*) = x_0(Q,R)$,则终止,子规划 $x_0(y^*)$ 的最优解便是原问题(P)的最优解.相反,置 $\Pi^{i+1} = \Pi^*$, $Q \cup \{\Pi^{i+1}\} \to Q$, $i+1 \to i$,返回(2).

例3 用交叉算法求解选址问题:

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij},$$

$$\text{s.t.}\ \sum_{i=1}^{n} x_{ij} = 1,\quad j = 1,2,\cdots,n,$$

$$\sum_{i=1}^{n} x_i = m,$$

$$x_{ij} - x_i \leqslant 0,\quad i,j = 1,2,\cdots,n;$$

$$x_i, x_{ij} = 0\ \text{或}\ 1,\quad i,j = 1,2,\cdots,n.$$

取拉格朗日松弛问题 R(Π) 为

$$\min\Big\{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}\Pi_j x_{ij} - \sum_{j=1}^{n}\Pi_j\Big\}$$

$$= \min\Big\{\sum_{i=1}^{n}\sum_{j=1}^{n}(c_{ij} + \Pi_j)x_{ij} - \sum_{j=1}^{n}\Pi_j\Big\},$$

$$\text{s.t.}\ \sum_{i=1}^{n} x_i = m,$$

$$x_{ij} - x_i \leqslant 0,$$

$$x_i, x_{ij} = 0\ \text{或}\ 1,\quad i,j = 1,2,\cdots,n.$$

将此问题分解:

取交叉算法中的整数向量 y 为0,1向量,$x^{\mathrm T} = (x_1, x_2, \cdots, x_n)$,使得

$$\sum_{i=1}^{n} x_i = m,$$

对给定的 x,子规划为

$$x_0(x) = \min\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij},$$

$$\text{s.t.}\ \sum_{i=1}^{n} x_{ij} = 1,\quad 0 \leqslant x_{ij} \leqslant x_i.$$

对偶子规划为

$$\max\left\{\sum_{j=1}^{n}\Pi_j - \sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}x_i\right\},$$

$$\text{s.t.}\ \Pi_j - W_{ij} \leqslant c_{ij},\quad i,j = 1,2,\cdots,n,$$

$$W_{ij} \geqslant 0,\qquad i,j = 1,2,\cdots,n.$$

当 x 是 0,1 向量时,子规划的最优解能使所有的 x_{ij} 取 0 或 1,因假如

$$x_1 = x_2 = \cdots = x_m = 1,$$

$$x_{m+1} = x_{m+2} = \cdots = x_n = 0,$$

设 $c_{i_S S} = \min\limits_{1\leqslant i\leqslant m} c_{is},\quad s = 1,2,\cdots,n$,则子规划的最优解为

$$x_{ij} = \begin{cases}1, & \text{当 } i = i_j \text{ 时;}\\ 0, & \text{当 } i \neq i_j \text{ 时.}\end{cases}\quad (i,j = 1,2,\cdots,n)$$

对偶子规划的最优解为

$$\Pi_j = c_{i_j j},\quad j = 1,2,\cdots,n,$$

$$W_{ij} = 0,\quad i = 1,2,\cdots,m,\quad j = 1,2,\cdots,n,$$

$$W_{ij} = \max\{0, c_{i_j j} - c_{ij}\},\quad i = m+1, m+2,\cdots,n,\quad j = 1,2,\cdots,n.$$

(1) 任取某 0-1,向量 $x^* = (x_1^*, x_2^*, \cdots, x_n^*)^{\mathrm{T}}$,使 $\sum_{i=1}^{n} x_i^* = m$.

(2) 解子规划 $x_0(x^*)$.

设子规划的最优解为 (x_{ij}^*),而对偶子规划的最优解为 (Π_j^*, W_{ij}^*).

(3) 求解拉格朗日松弛问题 R(Π),设其最优解为 $(\bar{x}_i^*, \bar{x}_{ij}^*)$.

(4) 若 $\bar{x}_i = x_i^*, i = 1,2,\cdots,n$,则终止,$(x_i^*, x_{ij}^*)$ 便是选址问题的最优解.相反,用 $\bar{x}_i^*$ 代替 $x_i^*, i = 1,2,\cdots,n$,返回(2).

4 背包问题的解法

背包问题是一个形式上最简单的 0-1 规划问题,然而也是最基本的整数规划问题.从理论上说,任何线性整数规划问题都可以化简成背包问题.

例如,若在整数规划的变量中,有某个整数变量 x_j,满足 $0 \leqslant x_j \leqslant 2^k$,则可通过变换

$$x_j = 2^k y_k + 2^{k-1} y_{k-1} + \cdots + 2y_1 + y_0,$$

$$y_0, y_1, \cdots, y_k \text{ 取 0 或 1},$$

化为0-1变量形式.而关于约束条件

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1,$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2,$$
$$x_i = 0 \text{或} 1, \quad i = 1,2,\cdots,n,$$

其中 a_{ij}, b_i 均为整数.设

$$\sum_{j=1}^{n} |a_{1j}| + |b_1| = d,$$

则易证:约束方程组解集和下述方程的解集完全相同,

$$\left(\sum_{j=1}^{n} a_{1j}x_j - b_1\right) + \lambda\left(\sum_{j=1}^{n} a_{2j}x_j - b_2\right) = 0,$$
$$x_j = 0 \text{或} 1, \quad j = 1,2,\cdots,n,$$

其中 λ 是大于 d 的整数.通过不断的两两合并就可将一般的线性整数问题化为背包问题.

4.1 动态规划解法

背包问题

$$\max\left\{\sum_{j=1}^{n} W_jx_j \mid \sum v_jx_j \leqslant v, x_j \text{取} 0 \text{或} 1\right\},$$

可归纳定义为如下子问题:

$$F_k(y) = \max\left\{\sum_{j=1}^{k} W_jx_j \mid \sum_{j=1}^{k} v_jx_j \leqslant y, x_j \text{取} 0 \text{或} 1\right\}.$$

其中 $k = 1,2,\cdots,n; y = 1,2,\cdots,v$.

$F_k(y) = 0$,当 $y \leqslant 0$ 时;

$F_0(y) = 0$,对任何整数 y.

容易验证:

1) 当 $y < v_k$ 时,$F_k(y) = F_{k-1}(y)$,

2) 当 $y > v_k$ 时,$F_k(y) = \max\{F_{k-1}(y), F_{k-1}(y - v_k) + W_k\}$,其中 $k = 2,3,\cdots,n; y = 1,2,\cdots,v$.显然,$F_n(v)$ 便是背包问题的解.

例1 设背包问题的数据如下:

$W_1 = 5, W_2 = 3, W_3 = 6, W_4 = 5, W_5 = 6,$

$v_1 = 6, v_2 = 3, v_3 = 7, v_4 = 4, v_5 = 5, v = 13.$

首先,易见

$F_1(y) = 0$,当 $y = 1,2,\cdots,5$ 时($x_1 = 0$);

$F_1(y) = 5$,当 $y = 6,7,\cdots,13$ 时($x_1 = 1$).

$F_2(y) = 0$,当 $y = 1,2$ 时($x_1 = x_2 = 0$);

$F_2(y) = 3$,当 $y = 3,4,5$ 时($x_1 = 0, x_2 = 1$).

然后,根据公式

$$F_k(y) = \max\{F_{k-1}(y), F_{k-1}(y - v_k) + w_k\},$$

可相继计算出

$F_2(y) = 5$，当 $y = 6,7,8$ 时（$x_1 = 1, x_2 = 0$）；

$F_2(y) = 8$，当 $y = 9,10,\cdots,13$ 时（$x_1 = x_2 = 1$）.

根据 $F_2(y)$ 的值，就可递推地算出 $F_3(y)$ 等. 整个过程如表 4-1 所示.

表 4-1

	W_1	W_2	W_3	W_4	W_5
	5	3	6	5	6
v	v_1	v_2	v_3	v_4	v_5
13	6	3	7	4	5
y	$F_1(y)$	$F_2(y)$	$F_3(y)$	$F_4(y)$	$F_5(y)$
1	0	0	0	0	0
2	0	0	0	0	0
3	0	3　$x_2 = 1$	3　$x_3 = 0$	3　$x_4 = 0$	3　$x_5 = 0$
4	0	3　$x_2 = 1$	3　$x_3 = 0$	5　$x_4 = 1$	5　$x_5 = 0$
5	0	3　$x_2 = 1$	3　$x_3 = 0$	5　$x_4 = 1$	6　$x_5 = 1$
6	5　$x_1 = 1$	5　$x_2 = 0$	5　$x_3 = 0$	5　$x_4 = 1$	6　$x_5 = 1$
7	5　$x_1 = 1$	5　$x_2 = 0$	6　$x_3 = 1$	8　$x_4 = 1$	8　$x_5 = 0$
8	5　$x_1 = 1$	5　$x_2 = 0$	6　$x_3 = 1$	8　$x_4 = 1$	9　$x_5 = 1$
9	5　$x_1 = 1$	8　$x_2 = 1$	8　$x_3 = 0$	8　$x_4 = 1$	11　$x_5 = 1$
10	5　$x_1 = 1$	8　$x_2 = 1$	9　$x_3 = 1$	10　$x_4 = 1$	11　$x_5 = 1$
11	5　$x_1 = 1$	8　$x_2 = 1$	9　$x_3 = 1$	11　$x_4 = 1$	11　$x_5 = 1$
12	5　$x_1 = 1$	8　$x_2 = 1$	9　$x_3 = 1$	11　$x_4 = 1$	14　$x_5 = 1$
13	5　$x_1 = 1$	8　$x_2 = 1$	11　$x_3 = 1$	13　$x_4 = 1$	14　$x_5 = 1$

最后得到此背包问题的最优解为

$$x_1^* = 0,\quad x_2^* = 1,\quad x_3^* = 0,\quad x_4^* = 1,\quad x_5^* = 1.$$

4.2 最短路方法

考虑如下的一般背包问题：

$$\min\left\{\sum_{j=1}^{n} c_j x_j \mid \sum_{j=1}^{n} a_j x_j = b, x_j \geqslant 0 \text{为整数}\right\},$$

其中 $c_j \geqslant 0$, a_j 和 b 都是正整数, $j = 1,2,\cdots,n$. 不妨设所有的 a_j 各不相同(因为假如 $a_k = a_j$, 且 $c_j \leqslant c_k$, 若问题有解, 则必有使 $x_k = 0$ 的最优解, 故可删去 x_k).

作一个有向图 $G = (V,E)$, V 为图的点集, E 为图的弧集, $V = \{0,1,2,\cdots,b\}$, 而

$$E = \{(h,i) \mid 0 \leqslant h < i \leqslant b, h, i \text{为整数, 且使} (i-h) \text{等于某个} a_j\},$$

若 $(h,i) \in E$, $a_j = i - h$, 则定义弧 (h,i) 的长度为 c_j.

根据上述定义, 易见求背包问题的最优解等价于求图 G 中点 0 到点 b 的定向最短路.

4.3 近似算法

考虑背包问题

$$\text{(KP)} \qquad \max\left\{\sum_{j=1}^{n} c_j x_j \mid \sum_{j=1}^{n} a_j x_j \leqslant a_0, x_j = 0 \text{或} 1\right\},$$

其中所有系数 c_j, a_j 都是正整数.

(KP) 的线性规划松弛问题

(LKP) 为

$$\max\left\{\sum_{j=1}^{n} c_j x_j \mid \sum_{j=1}^{n} a_j x_j \leqslant a_0, 0 \leqslant x_j \leqslant 1, 1 \leqslant j \leqslant n\right\}.$$

设变量已经过适当的排列, 使得

$$\frac{c_1}{a_1} \geqslant \frac{c_2}{a_2} \geqslant \cdots \geqslant \frac{c_n}{a_n},$$

则容易证明, (LKP) 的最优解必可取为如下的形式:

$$x_1 = x_2 = \cdots = x_{r-1} = 1, \quad x_r = \lambda,$$

$$x_{r+1} = x_{r+2} = \cdots = x_n = 0,$$

称 r 为问题(KP) 的界标.

对(KP) 的最优解 x^*, 定义

$$j' = \min\{j \mid x_j^* = 0, 1 \leqslant j \leqslant n\},$$

$$j'' = \max\{j \mid x_j^* = 1, 1 \leqslant j \leqslant n\},$$

$$j_1 = \min\{j', j''\},$$

$$j_2 = \max\{j', j''\},$$

$$G = \{j_1, j_1 + 1, \cdots, j_2\},$$

称 G 为(KP)的一个核,通常有 $j_1 = j', j_2 = j''$. 否则, $x^* = (1, \cdots, 1, 0, \cdots, 0)$. 假如(KP)的最优解不唯一,设 x^* 是使 $(j_2 - j_1)$ 最小的最优解. 称 $(j_2 - j_1)$ 为问题(KP)的核长. 显然, $r \in G$.

称子问题

$$\max\left\{\sum_{j\in G} c_j x_j \mid \sum_{j\in G} a_j x_j \leqslant a_0 - \sum_{j=1}^{j_1-1} a_j, x_j = 0 \text{ 或 } 1\right\}$$

为(KP)的核心问题,记作(GKP).

显然,求解(KP)和求解(GKP)等价.

巴拉斯(Balas)和泽梅尔(Zemel)作过统计试验,随机生成100个含有10 000个变量的背包问题,发现除少数问题外,核长都不超过25. 并且发现,只要当变量的数目足够大,核长的平均值与变量的数目无关. 后来,巴拉斯和泽梅尔从概率角度证明了以上事实,并由此提出了一个解大规模背包问题的近似算法. 它的基本解法如下:

(1) 排列变量,使满足

$$\frac{c_1}{a_2} \geqslant \frac{c_2}{a_2} \geqslant \cdots \geqslant \frac{c_n}{a_n}.$$

(2) 解(LKP),确定界标 r,设(LKP)的解为 $\bar{x}$.

(3) 选取一个适当大小的正整数 θ(例如取 $\theta = 12$).

(4) 定义

$$I = \{r-\theta, r-\theta+1, \cdots, r-1, r, r+1, \cdots, r+\theta\},$$
$$I_1 = \{r-\theta, r-\theta+1, \cdots, r-1\},$$
$$I_0 = \{r+1, r+2, \cdots, r+\theta\}.$$

(5) 定义

$$\bar{a} = \max_{k\in I_0} a_k, \quad \underline{a} = \min_{k\in I_0} a_k.$$

(6) 对 $i \in I_1$,定义

$$\bar{\omega}_i = \begin{cases} a_i + a_r\bar{x}_r, & \text{当 } a_i + a_r\bar{x}_r \leqslant \bar{a} \text{ 时;} \\ a_i - a_r(1-\bar{x}_r), & \text{当 } a_i + a_r\bar{x}_r > \bar{a} \text{ 时.} \end{cases}$$

$$\omega_i = \begin{cases} \bar{\omega}_i, & \text{当 } \bar{\omega}_i \geqslant \underline{a} \text{ 时;} \\ -\infty, & \text{当 } \bar{\omega}_i < \underline{a} \text{ 时.} \end{cases}$$

(7) 求解(KP)的计算程序:

1° 置 $i = r - \theta$,则有

$$x_j^* = \begin{cases} \bar{x}_j, & \text{当 } j \neq r \text{ 时;} \\ 0, & \text{当 } j = r \text{ 时.} \end{cases}$$

2° 若 $i = r$,则终止,此时的 x^* 便是所求的(近似最优)解. 若 $r > i$,则转3°.

3° 置 $\bar{x}^*$ 如下:

$$\bar{x}_j^* = \begin{cases} 0, & \text{当 } j = i \text{ 时;} \\ 1, & \text{当 } j = k \text{ 时;} \\ 0, & \text{当 } j = r, \text{而 } a_i + a_r\bar{x}_r \leqslant \bar{a} \text{ 时;} \\ 1, & \text{当 } j = r, \text{而 } a_i + a_r\bar{x}_r > \bar{a} \text{ 时;} \\ \bar{x}_j, & \text{当 } j \neq r, k, i \text{ 时.} \end{cases}$$

4° 若 $\sum_{j=1}^{n} c_j\bar{x}_j^* > \sum_{j=1}^{n} c_jx_j^*$，则用 $\bar{x}^*$ 替换 x^*，$i+1 \to i$，返回2°.

若 $\sum_{j=1}^{n} c_j\bar{x}_j^* \leqslant \sum_{j=1}^{n} c_jx_j^*$，则 $i+1 \to i$，转2°.

可以证明，上述程序能求得(KP)的最优解的概率，当变量数目趋于无穷时，概率趋向于1.

5　指派问题解法——匈牙利法

指派问题的模型为

$$\min z = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij},$$

$$\text{s.t.} \sum_{j=1}^{n} x_{ij} = 1, \quad i = 1,2,\cdots,n;$$

$$\sum_{i=1}^{n} x_{ij} = 1, \quad j = 1,2,\cdots,n,$$

$$x_{ij} = 0 \text{ 或 } 1.$$

它是一种特殊的0-1规划，它的一种简便的解法是由匈牙利数学家狄·考尼格(D. Koning)提出并证明了该方法中的主要定理的，故称匈牙利法.

匈牙利法的主要思路和步骤如下(此法把由 (c_{ij}) 所构成的矩阵称为费用矩阵)：

(1) 在费用矩阵中，任一行(列)减去或加上一个常数，其最优解集不变，只改变最优值.

(2) 用上述方法变换，使费用矩阵每行、每列都至少出现一个0，当能达到完全分配时，可令某些零元素所对应的变量 $x_{ij} = 1$(当然，分配时必须使每行、每列有且仅有一个元素 x_{ij} 为1). 于是可获得修改后的费用函数值为0，这必是此次的最优分配，否则只会使 $z \geqslant 0$.

例1　设指派问题中的费用矩阵 (c_{ij}) 为

$$(c_{ij}) = \begin{bmatrix} 2 & 10 & 9 & 7 \\ 15 & 4 & 14 & 8 \\ 13 & 14 & 16 & 11 \\ 4 & 15 & 13 & 9 \end{bmatrix},$$

(1) 矩阵(c_{ij})每行(每列)都减去该行(列)的最小元素以后,每行(列)至少出现一个0.

$$\begin{bmatrix} 2 & 10 & 9 & 7 \\ 15 & 4 & 14 & 8 \\ 13 & 14 & 16 & 11 \\ 4 & 15 & 13 & 9 \end{bmatrix} \xrightarrow[\text{分别减}2,4,11,4]{1\sim4\text{行}} \begin{bmatrix} 0 & 8 & 7 & 5 \\ 11 & 0 & 10 & 4 \\ 2 & 3 & 5 & 0 \\ 0 & 11 & 9 & 5 \end{bmatrix}$$

$$\xrightarrow{3\text{列减}5} \begin{bmatrix} 0 & 8 & 2 & 5 \\ 11 & 0 & 5 & 4 \\ 2 & 3 & 0 & 0 \\ 0 & 11 & 4 & 5 \end{bmatrix}.$$

(2) 试图制订一个完全分配方案,该方案只与表中零元素相对应.从第一行开始,依次检查各行,直到找出只有一个未标记的零元素的一行为止.如果在零元素上有符号△或×,则称零元素已标记,符号△表示相应的变量x_{ij}取1.对未做标记的零元素标△后,应对同一列的其他零元素画×.重复这一过程,直到每一行中没有尚未标记的零元素或至少有2个以上的零元素.本例中,因为第一行只有一个零元素,故标记符号△,并对第4行第1列的零元素画×;在第2行第2列标△;第3行有2个未标记的零元素,第4行没有未标记的零元素.至此,结果为

$$\begin{bmatrix} \overset{\triangle}{0} & 8 & 2 & 5 \\ 11 & \overset{\triangle}{0} & 5 & 4 \\ 2 & 3 & 0 & 0 \\ \not 0 & 11 & 4 & 5 \end{bmatrix}.$$

现在,依次检查每列,给每列只含一个未标记的零元素标△,对同一行的其他零元素画×(如果有的话).重复上述过程,直到每列中没有尚未标记的零元素或至少有2个零元素.本例中,第3列只有一个零元素,所以标上△,对第4列第3行的零元素画×,可得

$$\begin{bmatrix} \overset{\triangle}{0} & 8 & 2 & 5 \\ 11 & \overset{\triangle}{0} & 5 & 4 \\ 2 & 3 & \overset{\triangle}{0} & \not 0 \\ \not 0 & 11 & 4 & 5 \end{bmatrix}.$$

如果有多行多列同时有2个或2个以上的未标记的零元素,则可将其中的任意行或列中的一个未标记的零元素标△,并将同行和同列的其他零元素画×.重复上述步骤,直至所有零元素全部标完为止.

本例不可能制定出只含零元素的完全分配方案,于是按下列步骤画出最少数目的水平线和垂直线,以穿过所有零元素.

1° 检查尚未标记△的行,并记上√,得

$$
\begin{bmatrix}
\triangle\!\!\!\!\!\!0 & 8 & 2 & 5 \\
11 & \triangle\!\!\!\!\!\!0 & 5 & 4 \\
2 & 3 & \triangle\!\!\!\!\!\!0 & \not 0 \\
\not 0 & 11 & 4 & 5
\end{bmatrix}.
\begin{array}{l} \\ \\ \\ \checkmark \end{array}
$$

2° 在已打 ✓ 的行中对有零元素的列打 ✓.

3° 对在已打 ✓ 的列中对有标记 △ 的行打 ✓,其结果为

$$
\begin{array}{c}
\begin{bmatrix}
\triangle\!\!\!\!\!\!0 & 8 & 2 & 5 \\
11 & \triangle\!\!\!\!\!\!0 & 5 & 4 \\
2 & 3 & \triangle\!\!\!\!\!\!0 & \not 0 \\
\not 0 & 11 & 4 & 5
\end{bmatrix}.
\begin{array}{l} \checkmark \\ \\ \\ \checkmark \end{array} \\
\begin{array}{llll} \checkmark & \quad & \quad & \quad \end{array}
\end{array}
$$

4° 重复 2°,3° 直至不能再打 ✓ 为止.

5° 对未打 ✓ 的行和已打 ✓ 的列画线,即得满足要求的覆盖所有零元素且数目最少的水平线和垂直线.本例结果为

$$
\begin{array}{c}
\begin{bmatrix}
\triangle\!\!\!\!\!\!0 & 8 & 2 & 5 \\
11 & \triangle\!\!\!\!\!\!0 & 5 & 4 \\
2 & 3 & \triangle\!\!\!\!\!\!0 & \not 0 \\
\not 0 & 11 & 4 & 5
\end{bmatrix}.
\begin{array}{l} \checkmark \\ \\ \\ \checkmark \end{array} \\
\begin{array}{llll} \checkmark & \quad & \quad & \quad \end{array}
\end{array}
$$

在所有没有线通过的元素中找出最小元素,如本例中的 2.在未画线的行(如本例第 1,4 行)中每个元素均减去最小元素,得

$$
\begin{bmatrix}
-2 & 6 & 0 & 3 \\
11 & 0 & 5 & 4 \\
2 & 3 & 0 & 0 \\
-2 & 9 & 2 & 3
\end{bmatrix}.
$$

把出现负数的列再加上常数使负数成为 0,如本例第 1 列再加 2,得

$$
\begin{bmatrix}
0 & 6 & 0 & 3 \\
13 & 0 & 5 & 4 \\
4 & 3 & 0 & 0 \\
0 & 9 & 2 & 3
\end{bmatrix}.
$$

如此反复进行,直至能做出完全分配为止.本例至此已能做出完全分配:

$$\begin{bmatrix} \not{0} & 6 & \overset{\triangle}{0} & 3 \\ 13 & \overset{\triangle}{0} & 5 & 4 \\ 4 & 3 & \not{0} & \overset{\triangle}{0} \\ \overset{\triangle}{0} & 9 & 2 & 3 \end{bmatrix}.$$

本例的最优解为 $x_{13} = x_{22} = x_{34} = x_{41} = 1$,其余为 0,最优值即为整个过程中所减数之和减所加数之和,即 $z = 2 + 4 + 11 + 4 + 5 + 2 + 2 - 2 = 28$,也可由 $a_{13} + a_{22} + a_{34} + a_{41} = 28$ 得到.

6　集合覆盖问题解法

设 $I = \{1,2,\cdots,m\}$ 是一有限个元素的集合.设 $F = \{F_1,F_2,\cdots,F_n\}$,其中 F_j 是 I 的子集.称 F 是 I 上的一子集簇.记 $J = \{1,2,\cdots,n\}$. J 的一个子集 J^*,若满足

$$\bigcup_{j\in J^*} F_j = I,$$

则称 J^* 是 I 的一个覆盖.若 J^* 满足

$$F_j \cap F_k = \varnothing,\quad j \neq k, J, k \in J^*,$$

则称 J^* 是 F 的一个无关子簇.若 J^* 满足

$$F_j \cap F_k = \varnothing,\quad j \neq k, J, K \in J^*,$$

且

$$\bigcup_{j\in J^*} F_j = I,$$

则称 J^* 是 I 的一个分解.

定义　0,1 矩阵 $A = [a_{ij}]$ 为

$$a_{ij} = \begin{cases} 1, & \text{当元素 } i \in F_j \text{ 时}; \\ 0, & \text{当元素 } i \bar{\in} F_j \text{ 时}, \end{cases}$$

则称 A 为子集簇 F 的关联矩阵.设 c_j 是 F_j 的价格.定义 $J^*(J^* \leqslant J)$ 的价格为 $\sum\limits_{j\in J^*} c_j$,则有如下三个最基本的 0-1 规划问题:

1° 最优覆盖问题

$$\min x_0 = \sum_{j=1}^{n} c_j x_j,$$

$$\text{s.t.} \sum_{j=1}^{n} a_{ij} x_j \geqslant 1,\quad i = 1,2,\cdots,m,$$

$$x_j = 0 \text{ 或 } 1.$$

2° 最优分解问题

$$\min x_0 = \sum_{j=1}^{n} c_j x_j,$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = 1, \quad i = 1,2,\cdots,m,$$

$$x_j = 0 \text{ 或 } 1.$$

3° 最优无关子簇问题

$$\max x_0 = \sum_{j=1}^{n} c_j x_j,$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j \leqslant 1, \quad i = 1,2,\cdots,m,$$

$$x_j = 0 \text{ 或 } 1.$$

很多实际问题可以归结为上述问题.例如,设 $I = \{1,2,\cdots,m\}$ 表示某天运输任务的集合;$F_j(\subset I)$ 表示任务的某种搭配方式,它组合成一个合理的循环运输路线,而 F 表示所有合理的搭配方式的集合;设 c_j 是循环运输路线 F_j 上空驶的总里程,则此运输任务的分配问题就可归结为最优分解问题.又如,设 I 是材料的集合,J 是商品的种类,F_j 表示制作商品 j 时所必需的材料子集合(假设一种材料只能用在一种商品上,不能分开使用),c_j 表示商品 j 的价格,则此生产计划问题可以归结为无关子簇问题.

分解问题

$$\min\{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{1}, \boldsymbol{x} \text{ 是 } 0,1 \text{ 向量}\},$$

当 $\boldsymbol{c} > \boldsymbol{0}$ 时,可以化成如下等价的覆盖问题:

$$\min \sum_{j=1}^{n} (c_j + Lt_j) x_j,$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j \geqslant 1, \quad i = 1,2,\cdots,m,$$

$$x_j = 0 \text{ 或 } 1.$$

其中 $t_j = \sum_{i=1}^{m} a_{ij}, L = \sum_{j=1}^{n} c_j + 1,$

因此,下面只讨论覆盖问题的解法.

1. 问题化简

对覆盖问题(P),有

$$\min\{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \geqslant \boldsymbol{1}, \boldsymbol{x} \text{ 是 } 0,1 \text{ 向量}\},$$

若 $\boldsymbol{c} > \boldsymbol{0}$,那么常常可能根据以下的规则,预先确定某些 x_j 的值,或者去掉某些多余的约束条件.

1° 化简规则 1　若 $\boldsymbol{A}$ 的某一行 $\boldsymbol{a}_i$ 是一个单位向量,即 $a_{ik} = 1, a_{ij} = 0(\forall j \neq k)$,则 x_k 必须取 1.由于 $x_k = 1$,则在 $\boldsymbol{A}$ 的第 k 列中,凡是元素为 1 的行,条件都已得到满足,因此,可将这些行和第 k 列去掉.

2° 化简规则 2　若 A 中的行 a_t 和 a_p，使得 $a_t \geqslant a_p$，则第 t 个条件可以去掉，因为它是第 p 个条件的自然结果.

3° 化简规则 3　若有 A 的某列指标集 S，以及某一列指标 $k \bar{\in} S$，使得

$$\sum_{j \in S} a_{ij} \geqslant a_{ik} \quad i = 1,2,\cdots,m\ ,$$

$$\sum_{j \in S} c_j \leqslant c_k,$$

则显然第 k 列可以去掉，即 x_k 可取为 0.

2. 基本覆盖

对 A 的任一列指标子集合 J，定义并联向量 $X(J) = (x_1, x_2, \cdots, x_n)$ 如下：

$$x_j = \begin{cases} 1, & j \in J; \\ 0, & j \bar{\in} J, \end{cases}$$

若使 $X(J)$ 是覆盖问题的一个可行解，则称 J 为一个覆盖. 假如在覆盖 J 中，有一个指标 j^*，使 $J \setminus \{j^*\}$ 仍是一个覆盖，即

$$\sum_{j \in J} a_{ij} - a_{ij}^* \geqslant 1, \quad i = 1,2,\cdots,m,$$

则称 j^* 是过剩指标，一个不含过剩指标的覆盖，称为基本覆盖，否则称为过剩覆盖. 覆盖 J 中的一个指标 j'，它不是过剩的充要条件为

$$I(j') = \{i \mid \sum_{j \in J} a_{ij} - a'_{ij} = 0\} \neq \varnothing.$$

因为假定 $c > 0$，故覆盖问题的最优解必对应于一个基本覆盖.

记覆盖问题为(P)，记它的线性规划松弛问题为($\tilde{\mathrm{P}}$). 对($\tilde{\mathrm{P}}$)的任何一可行解 $\tilde{x}$，定义向量 $\bar{x}$ 如下：

$$\bar{x}_j = \min\{1, \tilde{x}_j\}, \quad j = 1,2,\cdots,n.$$

由于 A 是 0,1 矩阵，$\bar{x}$ 也是($\tilde{\mathrm{P}}$)的可行解. 设 $\tilde{x}^*$ 是($\tilde{\mathrm{P}}$)的一个基本最优解，因 $c > \mathbf{0}$，故必有

$$0 \leqslant \tilde{x}_j^* \leqslant 1, \quad j = 1,2,\cdots,n.$$

定义向量 $\bar{x}^*$ 如下：

$$\bar{x}_j^* = [\tilde{x}_j^*], \quad j = 1,2,\cdots,n,$$

显然 $\bar{x}^*$ 是覆盖问题(P)的一个可行解.

记

$$\bar{J}^* = \{j \mid \bar{x}_j^* = 1\},$$

则 $\bar{J}^*$ 是一个覆盖，但不一定是基本覆盖. 从 $\bar{J}^*$ 中逐个地减去过剩指标后，必可得到一个基本覆盖 J^*. 当然，用不同方式减过剩指标，可能获得不同的基本覆盖.

关于基本覆盖的重要性质：

1° 若 J 是一个基本覆盖，则 $X(J)$ 是($\tilde{\mathrm{P}}$)的一个基本可行解，经过行和列的适当排列后，矩阵 A 和向量 c 必可表示成如下的形式：

$$\begin{array}{ccc} & J\text{中的列} & J\text{以外的列} \\ c^{*T} = (& c_J, & c_N), \end{array}$$

$$A = \begin{bmatrix} I_1, & A_{12}, \\ A_{21}, & A_{22} \end{bmatrix}.$$

其中 I_1 是单位矩阵.

记

$$A_{12} = [p_1, p_2, \cdots, p_N], \quad c_N = (c'_1, c'_2, \cdots, c'_N),$$

则有

2° 若满足条件

$$c_J p_j - c'_j \leqslant 0, \quad j = 1, 2, \cdots, N,$$

则 $X(J)$ 是$(\tilde{P})$的基本最优解,因此 J 是(P)的一个最优覆盖.

3. 覆盖问题的割平面算法

对基本覆盖 J,若不满足性质 2°,则记

$$Q = \{j \mid c_j p_j - c'_j > 0, 1 \leqslant j \leqslant N\},$$

这时,若有另一个基本覆盖 J^*,使得 $c^T X(J^*) < c^T X(J)$,则 $X(J^*) = (x_J^*, x_N^*)$ 必满足 $\sum_{j \in Q} x_j^* \geqslant 1$. 称条件

$$\sum_{j \in Q} x_j \geqslant 1$$

为关于基本覆盖 J 的一个割平面.

算法程序:

(1) 若 A 中有一行全为 0,则终止,问题无可行解;相反,任给一覆盖,作为初始的记录 J,置 $x_0^* = c^T X(J)$.

(2) 利用化简规则 1,2,3,尽可能地化简问题(P)和矩阵 A,假如化简后已能完全确定(P)的最优解,则终止;否则,转(3)。

(3) 求松弛问题$(\tilde{P}$ 的基本最优解 $\tilde{x}^*$,并通过 $\tilde{x}^*$ 求得一个基本覆盖 J^*. 转(4).

(4) 若 $c^T X(J^*) < x_0^*$,则改进记录解,将 $J^* \to J$, $c^T X(J^*) \to x_0^*$,转(5).

(5) 将 A 和 c 排列成如下形式:

$$\begin{array}{ccc} & J^*\text{的列} & J^*\text{以外的列} \\ c^T = (& c_{J^*} & c_{N^*}), \end{array}$$

$$A = \begin{bmatrix} I_1 & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

记

$$A_{12} = [p_1, p_2, \cdots, p_{N^*}],$$

$$c_{N^*} = (c_1^*, c_2^*, \cdots, c_{N^*}^*),$$

转(6).

(6) 若满足

$$c_j^* p_j - c_j^* \leqslant 0, \quad j = 1,2,\cdots,N^*,$$

则终止，当时的记录解 $X(J)$ 便是(P) 的最优解；相反，置

$$Q = \{J \mid c_j^* p_j - c_j^* > 0 \mid 1 \leqslant j \leqslant N^*\},$$

转(7).

(7) 增加割平面条件

$$\sum_{j\in Q} x_j \geqslant 1.$$

用问题

$$\min\{c^{\mathrm{T}}x \mid Ax \geqslant 1, \quad \sum_{j\in Q} x_j \geqslant 1, x \text{ 为 } 0,1 \text{ 向量}\}$$

代替原来的问题(P)，返回(2).

7 非线性整数规划

对于非线性整数规划，目前尚无一种成熟而准确的求解方法．已有的方法，或是将线性整数规划的一些思路套用过来，或是针对一些特殊形式的非线性整数规划，因此都有很大的局限性．近来，一些将遗传算法应用于非线性整数规划的工作，取得了很好的效果．

7.1 字典序枚举法

在第4章中已叙述过任一有界整变量都能改用0-1变量来描述，而任一定义在0,1向量集合上的函数 $g(x)$，均能分解为两个关于每个变量 x_i 单调非增函数之差，因任一 0,1 向量 x，由方程

$$\prod_{j\in T} x_j \prod_{j\in J\setminus T} (1 - x_j) = 1$$

唯一确定，其中 $T \subseteq J = \{1,2,\cdots,n\}$．设 J^* 是 J 的子集的集合，于是

$$g(x) = \sum_{T\in J^*} \gamma(x(T)) \prod_{j\in T} x_j \prod_{j\in J\setminus T} (1 - x_j),$$

而对于任一目标函数 max $z(x)$，若 $z(x)$ 不是对每个变量 x_0 单调非增，则可修改成：加一约束 $-x_0 - z(x_0) \leqslant 0$ 和求 $\max(-x_0)$.

综上所述，任一有界整变量的整数规划都能化为形式：

$$\text{(P)} \quad \begin{aligned} &\max\ z(x), \\ &\text{s.t. } g_{i_1}(x) - g_{i_2}(x) \leqslant 0, \quad i = 1,2,\cdots,m, \\ &\qquad x \text{ 为 } n, 0,1 \text{ 维向量}, \end{aligned}$$

其中 $z(x), g_{i_1}(x), g_{i_2}(x)$ 均是关于每个变量 $x_i (i = 1,2,\cdots,n)$ 单调非增的.

针对上述形式就可运用以下的字典序枚举法：

(1) 若 $x=(0,0,\cdots,0)$ 是(P)的可行解,则它即为最优解;否则,设 $\underline{z}=-\infty$, $x=(0,0,\cdots,0,1)$,转(2).

(2) 若 $z(x)\leqslant \underline{z}$,转(5),其他则转(3).

(3) 若 x 是(P)的可行解,设 $\underline{z}=z(x)$,转(5);否则转(4).

(4) 若存在 x^* 和 i,使 $g_{i_1}(x^*-1)-g_{i_2}(x)>0$,转(5).其他,若 $x=(1,1,\cdots,1)$ 转(6);否则,令 $x=x+1$,返回(2).

(5) 设 $x=x^*$,返回(2).

(6) 终止.若 $\underline{z}=-\infty$,则(P)无解,其他,产生 $\underline{z}$ 的解,即为(P)的最优解.其中 x^*-1 和 $x+1$ 表示二元向量所相应的二进制数减1或加1.

7.2　拟布尔规划

把二元 n 维向量的简单实质函数称为**拟布尔函数**,任何拟布尔函数能表示为多项式

$$f(x)=\sum_{i=1}^{p}a_i\prod_{j=1}^{n}x_j^{k_{ij}},$$

其中 $k_{ij}=0$ 或1.

现考虑无约束拟布尔函数极大值问题

$$\max f(x),$$

把 $f(x)$ 再改写为

$$f(x)=x_1g_1(x_2,x_3,\cdots,x_n)+h_1(x_2,x_3,\cdots,x_n),$$

由此形式易见 $f(x)$ 的最优解中的分量 x_1^* 必满足

$$x_1^*(x_2,x_3,\cdots,x_n)=\begin{cases}1, & 若\ g_1(x_2,x_3,\cdots,x_n)\geqslant 0;\\ 0, & 若\ g_1(x_2,x_3,\cdots,x_n)<0.\end{cases}$$

于是,再把 $x_1^*(x_2,x_3,\cdots,x_n)$ 表示为用满足

$$g_1(x_2,x_3,\cdots,x_n)\geqslant 0$$

的变量 $x_2,x_3,\cdots,x_n$ 的每种组合枚举的多项式,而从 $f(x)$ 的表达式中消去变量 x_1.继续这个过程,到第 n 步时就使 $f(x)$ 的表达式中剩下一个变量,而用视察法就得再优 x_n^*,再递推而得 x_{n-1}^*,直至 x_1^* 为止.

7.3　蒙特卡罗法(随机取样法)

整数规划由于限制变量为整数而增加了难度,然而又由于整数解是有限个,于是为枚举法提供了方便.当然,在自变量维数很大和取值范围很宽的情况下,试图用显枚举法(即穷举法)计算出最优值是不现实的,但是应用概率理论可以证明,应用**蒙特卡罗法**经过一定的计算,完全可以得出一个满意解.

例1　设非线性整数规划

$$\max z=x_1^2+x_2^2+3x_3^2+4x_4^2+2x_5^2-8x_1-2x_2-3x_3-x_4-2x_5,$$

$$
\begin{aligned}
\text{s.t.}\quad & x_1 + x_2 + x_3 + x_4 + x_5 \leqslant 400,\\
& x_1 + 2x_2 + 2x_3 + x_4 + 6x_5 \leqslant 800,\\
& 2x_1 + x_2 + 6x_3 \leqslant 200,\\
& x_3 + x_4 + 5x_5 \leqslant 200,\\
& 0 \leqslant x_i \leqslant 99,\quad i = 1,2,\cdots,5.
\end{aligned}
$$

对此题,目前尚无有效方法求出准确解.如果用显枚举法试探,共需计算$(100)^5 = 10^{10}$个点,其计算量非常大.然而应用蒙特卡罗(Monte Carlo)法随机计算10^6个点,便可找到满意解.下面再应用概率理论估计一下可信度.

如果某整数规划用不同标本集描述目标函数的分布概率,其低值区的概率为0.999 99,高值区的概率仅为0.000 01.现计算10^6个点,则有任一个点能落在高值区的概率P_h应为

$$P_h = 1 - 0.999\,99^{1000000} \approx 0.999\,954\,602,$$

所以可信度也是相当高的.

当然,用计算机来计算10^6个标本点需很长时间,然而在微机日益普及的今天,这种矛盾日趋缓和.

7.4 罚函数 - 凑整算法

对于非线性混合整数规划,有

$$
\text{(P)}\qquad
\begin{aligned}
\min\ & z = f(\boldsymbol{x},\boldsymbol{y}),\\
\text{s.t.}\ & \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0},\\
& \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) \leqslant 0,\\
& \boldsymbol{x}^L \leqslant \boldsymbol{x} \leqslant \boldsymbol{x}^U,\ \boldsymbol{y}\ \text{为}\ 0,1\ \text{向量}.
\end{aligned}
$$

针对模型(P)的特点,可把0-1变量$\boldsymbol{y}$松弛为各分量在[0,1]上取值的连续变量,当其取值属于(0,1)时,则适当给以惩罚,在一定程度上迫使其靠近0或1取值.具体采用的一种措施是以惩罚函数$\phi(y)$(简称罚函数)取代目标函数$f(\boldsymbol{x},\boldsymbol{y})$中的0-1变量$\boldsymbol{y}$,即以$f(\boldsymbol{x},\varphi(\boldsymbol{y}))$代替目标函数$f(\boldsymbol{x},\boldsymbol{y})$,并把$\boldsymbol{y}$为0,1向量的约束松弛为$0 \leqslant y_i \leqslant 1$(对所有$i$),将(P)转化为如下的非线性规划:

$$
(\tilde{\text{x}})\qquad
\begin{aligned}
\min\ & f(\boldsymbol{x},\varphi(\boldsymbol{y})),\\
\text{s.t.}\ & \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = 0,\\
& \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) \leqslant 0,\\
& \boldsymbol{x}^L \leqslant \boldsymbol{x} \leqslant \boldsymbol{x}^U,\\
& 0 \leqslant y_j \leqslant 1,\ \text{对所有}\ j.
\end{aligned}
$$

其中罚函数例如可取$\varphi(y) = y^{1/3}$, $\varphi(y) = \frac{1}{\sin\omega}\sin\omega y$, ω可适当选取靠近但小于π的值,或采用其它形式.它们都具$\varphi(0) = 0$, $\varphi(1) = 1$,且$\varphi(y) \geqslant y$的特点,后者不仅在[0,1]上可导,而且当$y \in (0,1)$时,对目标函数施加的惩罚更强.当($\tilde{\text{P}}$)的解不符合整数约束时,可用凑整的方法进行处理.

7.5　相对差商法

考虑整数规划：

(P)
$$\min f(x),$$
$$\text{s.t. } g_j(x) \leqslant 0, \quad j = 1,2,\cdots,m,$$
$$x_i \in S_i, S_i \subseteq S, \quad i = 1,2,\cdots,n,$$

其中 S_i 为有限个离散值集合.

以下用 $x_{i,j}$ 表示变量 x_i 取 S_i 中第 j 个值.

当 $m = 1$ 时，称比值

$$\beta_i = \frac{\Delta g/\Delta x_i}{\Delta f/\Delta x_i} = \frac{\Delta g}{\Delta f} = \frac{g(x_{i,j+1}) - g(x_{i,j})}{f(x_{i,j+1}) - f(x_{i,j})},$$

为对应于设计变量 x_i 的**相对差商**.

当 $m > 1$ 时，首先进行归一化处理以克服因量纲不一致而出现约束值相差甚远的问题. 然后再采用统一约束的方法来处理.

在设计点 X 处，将约束函数集合 $G \xlongequal{\text{def}} \{g_j(X)\}$ 分为两个子集，$G_1 \xlongequal{\text{def}} \{g_i(X) \mid g_j(X) > 0\}$，$G_2 \xlongequal{\text{def}} \{g_j(X) \mid g_j(X) \leqslant 0\}$，定义

$$z(X) = \| G_1 \| \xlongequal{\text{def}} (\sum_{g_j \in G_1} g_j^2(X))^{1/2},$$

而相对差商定义为

$$\beta_i = \frac{\Delta z/\Delta x_i}{\Delta f/\Delta x_i} = \frac{\Delta z}{\Delta f} = \frac{z(x_{i,j+1}) - z(x_{i,j})}{f(x_{i,j+1}) - f(x_{i,j})},$$
$$\boldsymbol{B} = (\beta_1, \beta_2, \cdots, \beta_n),$$

称 $\boldsymbol{B}$ 为相对差商向量. 它的方向反映了约束函数相对于目标函数变化最快的方向.

当目标函数是设计变量的递增函数，约束函数是设计变量的递减函数时，可采用类似非线性规划确定搜索方向和步长的方法进行迭代，即采取迭代公式

$$\boldsymbol{X}^{(Q+1)} = \boldsymbol{X}^{(Q)} + \alpha \cdot \boldsymbol{D},$$

其中 $\boldsymbol{D}$ 为搜索方向，α 为步长. 由于整数规划及变量取值离散的特点，在 $\boldsymbol{D}$ 的方向上可能会没有设计点，以及步长 α 该由离散值之差来确定，因此将沿相对差商方向的搜索改为沿折线进行，正如在图 7-1 中所示的两个变量的情形，由 $P_1 \to P_2 \to P_3 \to P_4$.

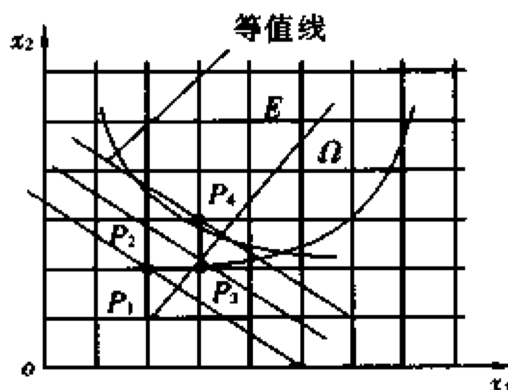


图 7-1

这样，迭代格式表示为

$$X^{(Q+1)} = X^{(Q)} + D \cdot \Delta X^{(Q)},$$

其中 $\Delta X^{(Q)}$ 为各设计变量下一离散值与当前离散值之差，而

$$D_i = \begin{cases} 1, & \beta_i = \min\{\beta_1, \beta_2, \cdots, \beta_n\}; \\ 0, & \beta_i \neq \min\{\beta_1, \beta_2, \cdots, \beta_n\}. \end{cases}$$

当优化过程接近最优点时，有时沿最小的方向前进会造成约束放松过多，而使目标函数不是最优的情况．这时可采用分支搜索的方法，即在计算过程中，若设计点 X_i 有 $z(X_i) = 0$，表示在该点约束条件全部满足，为一可行设计点；若有多个点 $i \in I, I = \{i \mid z(X_i) = 0\}$，则令 $f(X_j) = \min\limits_{i \in I} f(X_i)$，$\tilde{X} = X_j, f(\tilde{X}) = f(X_j)$，取 $\tilde{x}$ 为所求近似最优解．

综上分析，相对差商法程序为

(1) 将离散变量 $x_i (i = 1, 2, \cdots, n)$ 按目标函数的升序排列，形成离散变量集合 S_i．当在一定范围内目标函数和约束条件分别满足递增和递减时，集合 S_i 实际上就是离散变量的递增集合．

(2) 各设计变量 x_i 均取其在 S_i 中第一个离散值组成初始设计点，这个设计点通常不可行，但其目标函数值是最小的，$1 \times 10^{10} \to c$．

(3) 检查各约束条件，若约束条件全部满足，进行分支搜索，停止迭代，$\tilde{X}$ 即为近似最优解．

(4) 计算相对差商 β_i．

(5) 按迭代格式计算设计点 $x^{(Q+1)}$，返回(3)．

通过随机数值实验，说明本算法的计算精度是较高的，对结构优化设计中的非线性整数规划

$$\min f(x) = \sum_{i=1}^{n} c_i x_i,$$

$$\text{s.t.} \sum_{i=1}^{n} \frac{a_{ij}}{x_i + 1} \leqslant b_j, \quad j = 1, 2, \cdots, m,$$

$$x_i \in S_i, c_i \geqslant 0, a_{ij} \geqslant 0$$

是很有效的．

7.6 非线性 0-1 规划遗传算法的实现

非线性 0-1 规划

(P)
$$\min f(x, y),$$
$$\text{s.t. } g_i(x, y) \leqslant 0, \quad i = 1, 2, \cdots, m,$$
$$a \leqslant x \leqslant b,$$
$$x \in \mathbf{R}^n, y \text{ 为 } p \text{ 维 } 0,1 \text{ 向量}.$$

1. 适应值函数的取法

由于遗传算法为无约束最优化方法，约束问题需转化为无约束问题．例如，可

采用惩罚函数

$$h(x,y)=f(x,y)+\sum_i(r_i\varphi[g_i(x,y)]),$$

其中 r_i 为惩罚因子，

$$\varphi(t)=\begin{cases}0, & t\geqslant 0;\\ t^2, & t<0.\end{cases}$$

遗传算法是根据解点的适应值来判断好坏的，因此存在目标函数到适应值函数的对应问题.适应值函数的取法应遵循以下原则：

(1) 适应值函数应为非负.

(2) 目标函数的优化方向对应于适应值的增大方向.

例如，可取

$$\mathrm{fit}(x,y)=\begin{cases}\sup-h(x,y), & (x,y)\text{ 为可行点；}\\ [\sup-h(x,y)]u(k), & (x,y)\text{ 不是可行点.}\end{cases}$$

其中 $u(k)=1/(1+ck)$，$c>0$ 为调节因子，sup 为 $h(x,y)$ 的上确界.

2. 选择

选择的方式多种多样，例如，可采用较为简单的加权抽样法，其步骤为：计算出每一个体的适应值在总适应值中所占的比例 $L_i=f_i\Big/\sum f_i$，并据此将[0,1]分成 n 个小区间，每一小区间的长度为 L_i，利用随机数发生器产生一个在[0,1]上均匀分布的随机数 r，r 落入哪个小区间，相应的个体即被选中.

3. 编码

每个实数都与一个二进制数相对应，可取这个二进制数为该实数的编码.每个实数对应编码长为32位，前16位为整数部分，后16位为小数部分.对于多参数优化，其编码方法可用接排或错排.

4. 适应值调整

为提高遗传算法的收敛速度，以及避免提前收敛，适应值调整是有效手段之一.具体调整方法，例如可取

$$\mathrm{fit}'=\alpha(\mathrm{fit})+\beta,$$

fit 为原适应值，fit′ 为新适应值，α、β 应满足

$$\sum\mathrm{fit}'=\sum\mathrm{fit},$$

$$\mathrm{fit}'_{\min}=k(\mathrm{fit})_{\min},$$

其中 k 视种群内差距情况而定，若差距较大，则 $k>1$(如取1.618)；若差距较小，则 $k<1$(如取0.1).例如，也可取 $\mathrm{fit}'=\mathrm{fit}-\mathrm{fit}_{\min}$，其用在最小适应值也很大时，使适应值整体向下平移.

5. 变异

变异率例如可取 $0.005k$，其中 k 为进化代数.随着种群进化，变异率不断增大，其作用是增加种群的多样性，避免陷入局部极小.

6. 变量约束的处理

形如 $a\leqslant x\leqslant b$ 的约束称为变量界约束，将界约束中的上、下界改为形式 2^t，

其中 s 为非负整数，这样可以把变量界约束直接转换到编码中去，同时也避免了对一部分不可行区域的搜索，从而提高搜索效率.

7.0-1 变量的处理

对于 0-1 变量，其编码只有一位，任两个 0-1 变量交叉后，其值仍为 0 或 1，因此，对于 0-1 变量，只需以 0 或 1 作初始值，则整个过程中可以保证它始终为 0 或 1. 也可以采用将 0-1 变量连续化，再用惩罚函数加以处理的方法，但试算结果不如前者.

8. 结束准则

与传统优化方法不同，一般遗传算法没有明确的结束准则，往往是人为限定进化次数，达到次数即终止；或当某一类个体在种群中占到一定比例后，或连续几次最优值没有改进，便停止计算.

实践表明，用遗传算法求解复杂 0-1 规划是有效的.

遗传算法与传统算法不同之处为

1° 遗传算法搜索时使用的是编码，而不是变量本身.

2° 搜索过程是从一组点到另一组点，而传统算法是从一个点迭代到另一个点.

3° 淡化过程的随机性与传统算法的确定性搜索不同，遗传算法在适应性的控制下随机地选择、交叉、淘汰与变异，最后形成一个自组织、自适应的演化过程，因此，具有高度的自适应性.

4° 遗传算法简单通用，由于计算了多个点的值，能充分利用这些信息搜索全局最优. 其不足之处是计算量较大，缺乏较好的反馈准则，最优解与初始条件有关，参数选择困难等，有待继续工作.

参 考 文 献

1 许国志，马仲蕃著. 整数规划初步. 沈阳：辽宁教育出版社，1990.

2 Garfinkel R S, Nemhauser G L. Integer programming. New York: Wiley, 1972.

3 Tahe H A. Integer programming. New York: Academic Press, 1975.

4 孙焕纯，柴山，王跃方著. 离散变量结构优化设计. 大连：大连理工大学出版社，1995.

5 Nemhauser G L, Wolsey L A. Integer and combinatortial optimization. New York: Wiley, 1987.

·经济数学卷·

第10篇

动态规划

编　者　董加礼
审校者　胡毓达

目　　录

引　言

1951 年，美国数学家贝尔曼（R.Bellman）等根据一类所谓多阶段决策问题的特性，提出了解决这类问题的"最优化原理"，并研究了许多实际问题，从而创立了最优化的一个新分支——动态规划．1957 年贝尔曼出版了他的专著《动态规划》，这标志着动态规划理论的正式形成，并成为数学规划的一个重要分支．40 多年以来，动态规划在工程技术、经济领域和军事部门等众多方面都有重要应用，现今它已成为解决多阶段决策问题的一种有效方法．

根据决策过程是离散的还是连续的，是确定的还是随机的，动态规划大体上可以分为离散确定型、离散随机型、连续确定型和连续随机型等四种类型．这里将重点介绍前两类模型．

动态规划和其他的数学规划不同，它没有统一的数学模型，而对不同的问题要采用不同的方法去建立它们的模型．有了模型之后，要想得到数值解，仍然没有统一的处理方法．这是应当注意的．

1　动态规划原理

1.1　最短路问题及其解法

1.1.1　最短路问题及其特点

图 1-1 称为线路网络图，其中小圆圈称为点，两点间的连线称为弧，弧上的数字称为弧长．试求一条从起点 A 到终点 E 的连通弧，使其总弧长最短．称这类问题为最短路问题．

最短路问题的含义是广泛的，求解方法也有许多，下面介绍它的动态规划解法．

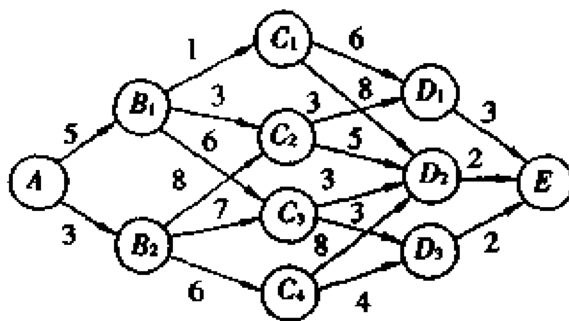


图 1-1

首先注意，从 A 到 E 的整个过程可以分成从 A 到 B，从 B 到 C、从 C 到 D，再从 D 到 E 四个阶段．每个阶段都有起点，如第二个阶段有两个起点 B_1 和 B_2，用 x_k 表示第 $k(k=1,2,3,4)$ 个阶段的起点，并称它为状态变量．从每个起点出发都有若干个选择，例如从 B_1 出发有三种选择，到 C_1 或到 C_2 或到 C_3，用 u_k 表示从第 $k(k=1,2,3,4)$ 个阶段的状态 x_k 出发

所作的选择,并称它为决策变量．如果用$f_k(x_k)$表示从第 k 个阶段的状态 x_k 出发到终点E 的最短弧长,或者用$f_k(x_k)$表示从起点 A 到第 k 个阶段的状态 x_k 的最短弧长,那么问题就变成求$f_1(x_1)=f_1(A)$,或者求 $f_5(x_5)=f_5(E)$.

其次,不难看出,如果最短路经过第 k 阶段的状态为 x_k,那么,从 x_k 出发到达终点E 的这条路线,对于从 x_k 出发到达终点E 的所有路线来说,显然也是最短路线．

根据最短路问题的上述特点,可有下述两种解法．

1.1.2 逆序解法

用$f_k(x_k)$表示从第 k 阶段的状态x_k 出发到终点E 的最短弧长,从后向前逐步求出各点到达终点 E 的最短路线的最短弧长,最后求出 $f_1(x_1)=f_1(A)$即为所求最短路线的最短弧长.计算步骤如下:

(1) 从最后一个阶段 $k=4$ 开始,按 f_4 的定义有

$$f_4(D_1)=3,\ f_4(D_2)=2,\ f_4(D_3)=2.$$

(2) 当 $k=3$ 时;因为第3阶段有4个状态,而每个状态又有两个决策可选取,所以有

$$f_3(C_1)=\min\begin{Bmatrix} d(C_1,D_1)+f_4(D_1) \\ d(C_1,D_2)+f_4(D_2) \end{Bmatrix}=\min\begin{Bmatrix} 6+3 \\ 8+2 \end{Bmatrix}=9,$$

其中 $d(\cdot,\cdot)$表示两点间的弧长．这说明从 C_1 到终点 E 的最短弧长为9,路径为 $C_1\to D_1\to E$,决策为 $u_3(C_1)=D_1$.

$$f_3(C_2)=\min\begin{Bmatrix} d(C_2,D_1)+f_4(D_1) \\ d(C_2,D_2)+f_4(D_2) \end{Bmatrix}=\min\begin{Bmatrix} 3+3 \\ 5+2 \end{Bmatrix}=6,$$

即从 C_2 到终点 E 的最短弧长为6,路径为 $C_2\to D_1\to E$,决策为 $u_3(C_2)=D_1$.

$$f_3(C_3)=\min\begin{Bmatrix} d(C_3,D_2)+f_4(D_2) \\ d(C_3,D_3)+f_4(D_3) \end{Bmatrix}=\min\begin{Bmatrix} 3+2 \\ 3+2 \end{Bmatrix}=5,$$

即从 C_3 到终点 E 的最短弧长为5,路径为 $C_3\to D_2$(或 D_3)$\to E$,决策为 $u_3(C_3)=D_2$(或 D_3).

$$f_3(C_4)=\min\begin{Bmatrix} d(C_4,D_2)+f_4(D_2) \\ d(C_4,D_3)+f_4(D_3) \end{Bmatrix}=\min\begin{Bmatrix} 8+2 \\ 4+2 \end{Bmatrix}=6,$$

即从 C_4 到终点 E 的最短弧长为6,路径为 $C_4\to D_3\to E$,决策为 $u_3(C_4)=D_3$.

(3) 当 $k=2$ 时,由于第2阶段有两个状态,每个状态又有3个决策可选,故有

$$f_2(B_1)=\min\begin{Bmatrix} d(B_1,C_1)+f_3(C_1) \\ d(B_1,C_2)+f_3(C_2) \\ d(B_1,C_3)+f_3(C_3) \end{Bmatrix}=\min\begin{Bmatrix} 1+9 \\ 3+6 \\ 6+5 \end{Bmatrix}=9,$$

即从 B_1 到终点 E 的最短弧长为9,路径为 $B_1\to C_2\to D_1\to E$.决策为 $u_2(B_1)=C_2$, $u_3(C_2)=D_1$, $u_4(D_1)=E$.

$$f_2(B_2)=\min\begin{Bmatrix} d(B_2,C_2)+f_3(C_2) \\ d(B_2,C_3)+f_3(C_3) \\ d(B_2,C_4)+f_3(C_4) \end{Bmatrix}=\min\begin{Bmatrix} 8+6 \\ 7+5 \\ 6+6 \end{Bmatrix}=12,$$

即从 B_2 到终点 E 的最短弧长为 12,路径为 $B_2\to C_3\to D_2$(或 D_3)$\to E$,或 $B_2\to C_4\to D_3\to E$;决策为 $u_2(B_2)=C_3$, $u_3(C_3)=D_2$(或 D_3), $u_4(D_2)=4$;或 $u_2(B_2)=C_4$, $u_3(C_4)=D_3$, $u_4(D_3)=E$.

(4) 当 $k=1$ 时,有

$$f_1(A)=\min\begin{Bmatrix} d(A,B_1)+f_2(B_1) \\ d(A,B_2)+f_2(B_2) \end{Bmatrix}=\min\begin{Bmatrix} 5+9 \\ 3+12 \end{Bmatrix}=14,$$

即从 A 到终点 E 的最短弧长为 14,路径为 $A\to B_1\to C_2\to D_1\to E$;决策为 $u_1(A)=B_1$, $u_2(B_1)=C_2$, $u_3(C_2)=D_1$, $u_4(D_1)=E$.

上述解法的四个步骤可归纳为下述递推公式:

$$\begin{cases} f_k(x_k)=\min\limits_{u_k\in D_k}\{d(x_k,x_{k+1})+f_{k+1}(x_{k+1})\}; \\ f_5(x_5)=0,\ k=4,3,2,1, \end{cases}$$

其中 $x_{k+1}=u_k(x_k)$,即从状态 x_k 出发,采取决策 u_k 到达下一状态 x_{k+1}; D_k 表示从状态 x_k 出发的所有可能选取的决策的集合;而 $f_5(x_5)=0$ 称为边界条件,因为状态 $x_5=E$ 已是终点.

这个递推公式就是最短路问题的数学模型,也叫动态规划方程.

由于这种算法的寻优方向与过程的行进方向刚好相反,故称逆序解法.

1.1.3　顺序解法

用 $f_k(x_k)$ 表示从起点 A 到第 k 阶段的状态 x_k 的最短弧长,从前向后逐步求出起点 A 到各阶段起点的最短弧长,最后也可求出从起点 A 到终点 E 的最短弧长及其对应的路径. 计算步骤如下:

按定义显然有 $f_1(x_1)=f_1(A)=0$,称它为边界条件. 以下从第二阶段 $k=2$ 开始计算.

(1) 当 $k=2$ 时,按 f_2 的定义有

$$f_2(B_1)=5,\ f_2(B_2)=3.$$

(2) 当 $k=3$ 时,按 f_3 的定义分别有

$$f_3(C_1)=d(B_1,C_1)+f_2(B_1)=1+5=6,$$

$$f_3(C_2)=\min\begin{Bmatrix} d(B_1,C_2)+f_2(B_1) \\ d(B_2,C_2)+f_2(B_2) \end{Bmatrix}=\min\begin{Bmatrix} 3+5 \\ 8+3 \end{Bmatrix}=8,$$

$$f_3(C_3)=\min\begin{Bmatrix} d(B_1,C_3)+f_2(B_1) \\ d(B_2,C_3)+f_2(B_2) \end{Bmatrix}=\min\begin{Bmatrix} 6+5 \\ 7+3 \end{Bmatrix}=10,$$

$$f_3(C_4)=d(B_2,C_4)+f_2(B_2)=6+3=9.$$

(3) 当 $k=4$ 时,按 f_4 的定义分别有

$$f_4(D_1)=\min\begin{Bmatrix}d(C_1,D_1)+f_3(C_1)\\d(C_2,D_1)+f_3(C_2)\end{Bmatrix}=\min\begin{Bmatrix}6+6\\3+8\end{Bmatrix}=11,$$

$$f_4(D_2)=\min\begin{Bmatrix}d(C_1,D_2)+f_3(C_1)\\d(C_2,D_2)+f_3(C_2)\\d(C_3,D_2)+f_3(C_3)\\d(C_4,D_2)+f_3(C_4)\end{Bmatrix}=\min\begin{Bmatrix}8+6\\5+8\\3+10\\8+9\end{Bmatrix}=13,$$

$$f_4(D_3)=\min\begin{Bmatrix}d(C_3,D_3)+f_3(C_3)\\d(C_4,D_3)+f_3(C_4)\end{Bmatrix}=\min\begin{Bmatrix}3+10\\4+9\end{Bmatrix}=13.$$

(4) 当 $k=5$ 时,按 f_5 的定义有

$$f_5(E)=\min\begin{Bmatrix}d(D_1,E)+f_4(D_1)\\d(D_2,E)+f_4(D_2)\\d(D_3,E)+f_4(D_3)\end{Bmatrix}=\min\begin{Bmatrix}3+11\\2+13\\2+13\end{Bmatrix}=14.$$

$f_5(E)=14$ 为所求的最短弧长,路径为 $A\to B_1\to C_2\to D_1\to E$,决策为 $u_1(A)=B_1$, $u_2(B_1)=C_2$, $u_3(C_2)=D_1$, $u_4(D_1)=E$,与逆序解法的结果完全一样.

上述解法也可写成统一的递推公式形式:

$$\begin{cases}f_k(x_k)=\min\limits_{u_{k-1}\in D_{k-1}}\{d(u_{k-1},x_k)+f_{k-1}(x_{k-1})\},\\f_1(x_1)=0,\quad k=2,3,4,5,\end{cases}$$

其中 $x_{k-1}=u_{k-1}(x_k)$,即从第 k 阶段的起点状态 x_k 通过 u_{k-1} 去寻找第 $k-1$ 阶段的起点 x_{k-1}, $f_1(x_1)=0$ 称为边界条件.

由于这种算法的寻优方向与过程的行进方向相同,故称顺序解法.

1.2 动态规划的基本概念和术语

1.2.1 多阶段决策问题

如果一个问题的整个过程可以分成若干个互相联系的阶段,每个阶段都需要作出决策,而当每个阶段的决策都确定之后,整个过程就确定了,那么这个问题就叫做**多阶段决策问题**.动态规划就是解决这类问题的一种重要方法.

一个问题是不是多阶段决策问题是需要判断的,而这种判断并没有一般规律可循,只能靠经验和技巧.

1.2.2 阶段变量

描述多阶段决策问题阶段数的变量叫**阶段变量**,记作 $k(k=1,2,\cdots)$.如果阶段变量是确定的、有限的,而且在决策前就知道其数值,则称此问题为**定期问题**.如果阶段变量虽然是确定的、有限的,但在决策之前并不知道它的数值,则称此问题为**不定期问题**.

1.2.3 状态变量

对于多阶段决策问题，每一阶段的起始"位置"叫状态，它既是该阶段的某一起点，也是前一阶段的终点．不同问题其状态的含义是不同的．

描述过程状态的变量叫**状态变量**，用 x_k 表示第 k 阶段的某一状态．

1.2.4 决策变量

将过程由一个状态变到另一个状态的决定或选择叫做**决策**．描述决策的变量叫**决策变量**，用 $u_k(x_k)$表示从第 k 阶段的状态 x_k 处所采取的决策．在第 k 阶段的状态 x_k 处的所有决策构成的集合叫做**决策集合**，记作 $D_k(x_k)$.

1.2.5 整体策略

对于 n 阶段决策问题，当每个阶段的决策都确定以后，由每个阶段的决策 $u_k(x_k)(k=1,2,\cdots,n)$所构成的决策序列称为一个**整体策略**，简称**策略**，记作 $p_{1,n}(x_1)$，即

$$p_{1,n}(x_1)=\{u_1(x_1),u_2(x_2),\cdots,u_n(x_n)\}.$$

而

$$p_{k,n}(x_k)=\{u_k(x_k),u_{k+1}(x_{k+1}),\cdots,u_n(x_n)\}$$

和

$$p_{1,k}(x_1)=\{u_1(x_1),u_2(x_2),\cdots,u_k(x_k)\}$$

则分别称为一个**后部 k 段子策略**和**前部 k 段子策略**．用 $P_{1,n}(x_1)$、$P_{1,k}(x_1)$及 $P_{k,n}(x_k)$分别表示**整体策略集合**，**前部及后部 k 段子策略集合**．

如果一个策略使得多阶段决策问题达到所要求的最优，则称此策略为**最优策略**．

1.2.6 状态转移方程

把过程由一个状态变到另一个状态叫做**状态转移**．显然它既与状态有关，又与决策有关．

如果第 k 阶段的状态 x_k 和决策 u_k 都确定以后，第 $k+1$ 阶段的状态 x_{k+1}就随之确定，那么就把这个对应关系记作

$$x_{k+1}=T_k(x_k,u_k),$$

并称它为由状态 x_k 到 x_{k+1}的**顺序状态转移方程**．

如果第 k 阶段的状态 x_k 和第 $k-1$ 阶段的决策 u_{k-1}都确定以后，第 $k-1$ 阶段的状态 x_{k-1}就随之确定，则把这个对应关系记作

$$x_{k-1}=T_{k-1}(u_{k-1},x_k),$$

并称它为由状态 x_k 到 x_{k-1}的**逆序状态转移方程**．

1.2.7 指标函数

每个多阶段决策问题都存在很多策略，而每个策略都会对应某种"效益"．不

同问题效益的含义是不同的,同一问题采取不同的策略其效益也会不一样. 衡量问题效益优劣的数量指标称为**指标函数**.

对 n 阶段决策问题,用 $F_{k,n}(x_k,p_{k,n})$ 表示从第 k 阶段的状态 x_k 出发,采用策略 $p_{k,n}$ 到达终点 x_{k+1} 的**后部指标函数**. 若上述过程采用的是最优策略 $p_{k,n}^*$,则相应的后部指标函数记作 $f_{k,n}(x_k,p_{k,n}^*)$,简记为 $f_k(x_k)$,并称为**后部最优指标函数**,简称为**最优函数**. $f_k(x_k)$ 与 $F_{k,n}(x_k,p_{k,n})$ 间的关系为

$$f_k(x_k)=F_{k,n}(x_k,p_{k,n}^*)=\underset{p_{k,n}\in P_{k,n}}{\mathrm{opt}}F_{k,n}(x_k,p_{k,n}),$$

这里 opt 是 optimization 的缩写,表示最优,通常取 max 或 min. 这时 $f_1(x_1)$ 表示**整体最优函数**.

类似地有**前部指标函数** $F_{1,k}(x_1,p_{1,k})$ 和**前部最优指标函数** $f_{1,k}(x_1,p_{1,k}^*)$,并同样记作 $f_k(x_k)$,且

$$f_k(x_k)=F_{1,k}(x_1,p_{1,k}^*)=\underset{p_{1,k}\in P_{1,k}}{\mathrm{opt}}F_{1,k}(x_1,p_{1,k}),$$

这时 $f_{n+1}(x_{n+1})$ 为**整体最优函数**.

用 $d(x_k,x_{k+1})$ 表示状态 x_k 与 x_{k+1} 之间对应的指标,称为**阶段指标**. 当过程是由状态 x_k 出发,采取决策 u_k 到达状态 $x_{k+1}=T_k(x_k,u_k)$ 时,则把 $d(x_k,x_{k+1})$ 写成 $d(x_k,u_k)$;当过程是由 x_{k+1} 出发,采取决策 u_k 去确定状态 $x_k=T_k(u_k,x_{k+1})$ 时,则把 $d(x_k,x_{k+1})$ 写成 $d(u_k,x_{k+1})$.

指标函数通常采用如下两种形式:

(1) 指标函数为阶段指标之和的形式,即

$$F_{k,n}=\sum_{j=k}^{n}d(x_j,u_j)=d(x_k,u_k)+F_{k+1,n},$$

$$F_{1,k}=\sum_{j=2}^{k}d(u_{j-1},x_j)=d(u_{k-1},x_k)+F_{1,k-1}.$$

(2) 指标函数为阶段指标之积的形式,即

$$F_{k,n}=\prod_{j=k}^{n}d(x_j,u_j)=d(x_k,u_k)\cdot F_{k+1,n},$$

$$F_{1,k}=\prod_{j=2}^{k}d(u_{j-1},x_j)=d(u_{k-1},x_k)\cdot F_{1,k-1}.$$

1.3　最优化原理与动态规划方程

1.3.1　最优化原理

对于多阶段决策问题,作为整个过程的最优策略必然具有这样的性质:无论过去的状态和决策如何,就所形成的状态而言,余下的诸策略必然构成一个最优子策略. 多阶段决策问题的这一规律称为**最优化原理**.

1.3.2 逆序动态规划方程

对后部指标函数 $F_{k,n}$ 及最优函数 $f_k(x_k)$ 有

(1) 当 $F_{k,n}=\sum_{j=k}^{n}d(x_j,u_j)$ 时，$f_k(x_k)$ 满足递推方程

$$\begin{cases}f_k(x_k)=\underset{u_k\in D_k}{\mathrm{opt}}\{d(x_k,u_k)+f_{k+1}(x_{k+1})\},\\ f_{n+1}(x_{n+1})=0,k=n,n-1,\cdots,2,1.\end{cases}$$

(2) 当 $F_{k,n}=\prod_{j=k}^{n}d(x_j,u_j)$ 时，$f_k(x_k)$ 满足递推方程

$$\begin{cases}f_k(x_k)=\underset{u_k\in D_k}{\mathrm{opt}}\{d(x_k,u_k)\cdot f_{k+1}(x_{k+1})\},\\ f_{n+1}(x_{n+1})=1,k=n,n-1,\cdots,2,1.\end{cases}$$

利用这两个递推公式原则上可求出最优函数 $f_1(x_1)$．称这两种递推公式为**逆序动态规划方程**．这种求最优函数的方法叫**逆序法**．

1.3.3 顺序动态规划方程

对前部指标函数 $F_{1,k}$ 及最优函数 $f_k(x_k)$ 有

(1) 当 $F_{1,k}=\sum_{j=2}^{k}d(u_{j-1},x_j)$ 时，$f_k(x_k)$ 满足递推方程

$$\begin{cases}f_k(x_k)=\underset{u_{k-1}\in D_{k-1}}{\mathrm{opt}}\{d(u_{k-1},x_k)+f_{k-1}(x_{k-1})\},\\ f_1(x_1)=0,k=2,3,\cdots,n+1.\end{cases}$$

(2) 当 $F_{1,k}=\prod_{j=2}^{k}d(u_{j-1},x_j)$ 时，$f_k(x_k)$ 满足递推公式

$$\begin{cases}f_k(x_k)=\underset{u_{k-1}\in D_{k-1}}{\mathrm{opt}}\{d(u_{k-1},x_k)\cdot f_{k-1}(x_{k-1})\},\\ f_1(x_1)=1,k=2,3,\cdots,n+1.\end{cases}$$

利用这两个递推公式原则上可以求出最优函数 $f_{n+1}(x_{n+1})$．称这两种递推公式为**顺序动态规划方程**．这种求最优函数的方法叫**顺序法**．

1.4 动态规划基本定理

基本定理 对于 n 阶段决策问题，若 $p_{1,n}^*$ 是最优策略，则对任意满足 $1<k<n$ 的自然数 k，其子策略 $p_{k,n}^*$（或 $p_{1,k}^*$）对于以

$$x_k=T_{k-1}(x_{k-1},u_{k-1}^*)\quad(\text{或 } x_{k-1}=T_{k-1}(u_{k-1}^*,x_k))$$

为初始状态的 k 到 n（或 1 到 k）段子过程来说，也必定是最优策略．

1.5　例题——生产与存储问题

例 1　假设某车间每月底都要供应总装车间一定数量的部件．由于生产条件的变化，该车间每月生产单位部件所耗费的工时不同；每月的生产量除供本月需要外，剩余部分可存入仓库供备用．现已知在半年内各月份的需求量及生产该部件每单位数所需工时，如表 1-1 所示．

表 1-1

月份 k	0	1	2	3	4	5	6
月需求量 b_k	0	8	5	3	2	7	4
单位工时 a_k	11	18	13	17	20	10	

设库存容量 $H=9$，开始时库存量为 2，期终库存量为 0．要求制定一个半年逐月生产计划，使得既满足需求和库存容量的限制，又使得耗费的总工时数为最少．

解　这是一个确定型的多阶段决策问题．按月份划分阶段，阶段变量 $k=0,1,\cdots,6$；状态变量 x_k 取第 k 月的部件库存量（上月产品送入后，本月需求量送出前）；决策变量 u_k 取第 k 月生产的部件数．于是状态转移方程及决策集合等分别为

$$x_{k+1}=x_k+u_k-b_k,$$
$$D_k(x_k)=\{u_k\geqslant 0\mid b_k\leqslant x_k\leqslant H\},$$
$$d(x_k,u_k)=a_ku_k,$$
$$F_{k,6}=\sum_{j=k}^{6}a_ju_j.$$

如果用 $f_k(x_k)$ 表示在状态 x_k 之下从第 k 月到 6 月末生产部件的最少累计工时数，则由最优化原理，便得描述此问题的逆序动态规划方程为

$$\begin{cases}f_k(x_k)=\min\limits_{u_k\in D_k}\{a_ku_k+f_{k+1}(x_k+u_k-b_k)\},\\ f_7(x_7)=0,\ k=6,5,4,3,2,1,0.\end{cases}$$

按逆序法从后往前计算：

因期终库存为 0，故 $x_7=0,u_6=0$；又 $b_6=4$，于是由状态转移方程 $x_7=x_6+u_6-b_6$，得 $x_6=4$，从而

$$f_6(x_6)=\min_{u_6\in D_6}(a_6u_6+0)=0,u_6^*=0.$$

又因 $4=b_6\leqslant x_5+u_5-b_5\leqslant H=9$，而 $b_5=7$，所以

$$0\leqslant 11-x_5\leqslant u_5\leqslant 16-x_5,$$

于是

$$f_5(x_5)=\min_{u_5\in D_5}\{a_5u_5+f_6(x_6)\}=\min_{11-x_5\leqslant u_5\leqslant 16-x_5}\{10u_5+0\}=110-10x_5,$$
$$u_5^*=11-x_5.$$

再由 $7=b_5\leqslant x_4+u_4-b_4\leqslant H=9$，而 $b_4=2$，得

$$9 - x_4 \leqslant u_4 \leqslant 11 - x_4,$$

所以

$$f_4(x_4) = \min_{u_4 \in D_4} \{a_4 u_4 + f_5(x_5)\} = \min_{9-x_4 \leqslant u_4 \leqslant 11-x_4} \{20u_4 + 110 - 10(x_4 + u_4 - 2)\}$$

$$= 220 - 20x_4, u_4^* = 9 - x_4.$$

类似可分别求出

$$f_3(x_3) = 244 - 17x_3, \quad u_3^* = 12 - x_3;$$

$$f_2(x_2) = 273 - 13x_2, \quad u_2^* = 14 - x_2;$$

$$f_1(x_1) = 442 - 18x_1, \quad u_1^* = 13 - x_1;$$

$$f_0(x_0) = 393 - 18x_0, \quad u_0^* = 7.$$

注意 $x_0 = 2$,故最少工时 $f_0(2) = 393 - 36 = 357$.

最后求最优策略:

$$x_0 = 2, \qquad u_0^* = 7;$$

$$x_1 = x_0 + u_0 - b_0 = 9, \quad u_1^* = 13 - x_1 = 4;$$

$$x_2 = x_1 + u_1 - b_1 = 5, \quad u_2^* = 14 - x_2 = 9;$$

$$x_3 = x_2 + u_2 - b_2 = 9, \quad u_3^* = 12 - x_3 = 3;$$

$$x_4 = x_3 + u_3 - b_3 = 9, \quad u_4^* = 9 - x_4 = 0;$$

$$x_5 = x_4 + u_4 - b_4 = 7, \quad u_5^* = 11 - x_5 = 4.$$

即最优逐月生产计划为

7, 4, 9, 3, 0, 4.

2 不定期多阶段决策问题的两种解法

设有 $n+1$ 个点:1,2,…,$n+1$. 任意两点 i 和 j 有一弧连接,其长度记为 C_{ij}:$0 \leqslant C_{ij} \leqslant \infty$, $C_{ij} = 0$ 表示 i 和 j 退缩为一点, $C_{ij} = \infty$ 表示 i 与 j 之间不存在连接它们的弧. 试求任意点 $i(i = 1,2,\cdots,n)$ 至固定点 $n+1$ 的最短路线.

由于点 i 至固定点 $n+1$ 要经过哪些点事先是不知道的,因此这是不定期问题. 因为这类问题的阶段数不定,故无法直接建立其动态规划方程. 下面给出两种解法:**函数空间迭代法和策略空间迭代法**.

2.1 函数空间迭代法

2.1.1 迭代程序

(1) 先选取初始函数 $f_1(i)$:

$$\begin{cases} f_1(i) = C_{i(n+1)}, i = 1,2,\cdots,n+1; \\ f_1(n+1) = 0, \end{cases} \tag{2-1}$$

这里 $f_1(i)$ 表示由 i 点出发向固定点 $n+1$ 行进一步的最短距离．

(2) 利用下述递推公式作函数列 $\{f_k(i)\}$：

$$\begin{cases} f_k(i) = \min\limits_j \{C_{ij} + f_{k-1}(j)\}, k = 2,3,\cdots; \\ f_k(n+1) = 0, i = 1,2,\cdots,n, \end{cases} \tag{2-2}$$

这里 $f_k(i)$ 表示从 i 点出发向固定点 $n+1$ 行进 k 步的最短距离，并规定向前行进时，已走过的点不再重复，以免发生回路．

(3) 若经过 $k+1$ 步，使得

$$f_{k+1}(i) = f_k(i)$$

对一切 $i = 1,2,\cdots,n$ 都成立，则停止迭代，$f_k(i)$ 就是从 i 点出发到固定点 $n+1$ 的最短距离，并且行进了 k 步．否则，转步(2) 继续迭代．

2.1.2 收敛性定理

定理 1 由迭代公式(2-1) 和(2-2) 确定的函数列 $\{f_k(i)\}$ 满足

1° $\{f_k(i)\}$ 单调下降，且收敛于函数方程

$$\begin{cases} f(i) = \min\limits_j \{C_{ij} + f(j)\}, \\ f(n+1) = 0, i = 1,2,\cdots,n \end{cases} \tag{2-3}$$

的解 $f(i)$，这里 $f(i)$ 表示由点 i 出发到固定点 $n+1$ 的最短距离；

2° $\{f_k(i)\}$ 在 n 步内收敛于 $f(i)$．

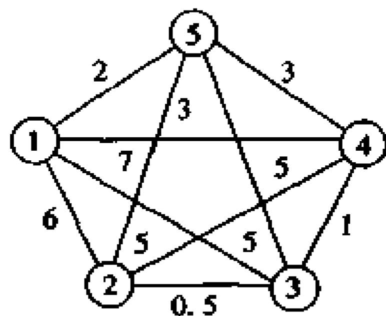


图 2-1

例 1 设有1,2,3,4,5五个城市，城市与城市间连线上的数字表示二者的距离，如图 2-1 所示．试求各城市到 5 城的最短距离及最短路线．

解 由(2-1) 式及图 2-1 知，

$$f_1(1) = C_{15} = 2, f_1(2) = C_{25} = 7,$$
$$f_1(3) = C_{35} = 5, f_1(4) = C_{45} = 3.$$

当 $k = 2$ 时，在公式

$$f_2(i) = \min\{C_{i1} + f_1(1), C_{i2} + f_1(2), C_{i3} + f_1(3), C_{i4} + f_1(4), C_{i5} + f_1(5)\}$$

中，令 $i = 1,2,3,4$，得

$$f_2(1) = \min\{0+2, 6+7, 5+5, 3+3, 2+0\} = 2,$$
$$f_2(2) = \min\{6+2, 0+7, 0.5+5, 5+3, 7+0\} = 5.5,$$
$$f_2(3) = \min\{5+2, 0.5+7, 0+5, 1+3, 5+0\} = 4,$$
$$f_2(4) = \min\{3+2, 5+7, 1+5, 0+3, 3+0\} = 3.$$

$f_2(i) \neq f_1(i)$，继续迭代．

当 $k = 3$ 时，在公式

$$f_3(i) = \min\{C_{i1} + f_2(1), C_{i2} + f_2(2), C_{i3} + f_2(3), C_{i4} + f_2(4), C_{i5} + f_2(5)\}$$

中，令 $i = 1,2,3,4$，得

$$f_3(1) = \min\{0+2, 6+5.5, 5+4, 3+3, 2+0\} = 2,$$
$$f_3(2) = \min\{6+2, 0+5.5, 0.5+4, 5+3, 7+0\} = 4.5,$$
$$f_3(3) = \min\{5+2, 0.5+5.5, 0+4, 1+3, 5+0\} = 4,$$
$$f_3(4) = \min\{3+2, 5+5.5, 1+4, 0+3, 3+0\} = 3.$$

$f_3(i) \neq f_2(i)$,继续迭代.

当 $k=4$ 时,在公式

$$f_4(i) = \min\{C_{i1}+f_3(1), C_{i2}+f_3(2), C_{i3}+f_3(3), C_{i4}+f_3(4), C_{i5}+f_3(5)\}$$

中,令 $i=1,2,3,4$,得

$$f_4(1) = \min\{0+2, 6+4.5, 5+4, 3+3, 2+0\} = 2,$$
$$f_4(2) = \min\{6+2, 0+4.5, 0.5+4, 5+3, 7+0\} = 4.5,$$
$$f_4(3) = \min\{5+2, 0.5+4.5, 0+4, 1+3, 5+0\} = 4,$$
$$f_4(4) = \min\{3+2, 5+4.5, 1+4, 0+3, 3+0\} = 3.$$

因为对 $i=1,2,3,4$ 均有 $f_4(i)=f_3(i)$,故停止迭代,1,2,3,4 城到 5 城的最短距离分别为 2,4.5,4,3.

为求出最短路线,须利用 $f_4(i)$ 的计算过程,反向追踪,找出最优决策 $u^*(i)=j^*$,即若

$$f_4(i) = \min_{1\leq j\leq 5}\{C_{ij}+f_3(i)\} = C_{ij^*}+f_3(j^*),$$

则 $u^*(i)=j^*$.但不能取含有 $C_{ij}=0$ 的城作为 $u^*(i)$,因为那样的路线不能说明是最优的.据此可得

$$u^*(1)=5, u^*(2)=3, u^*(3)=4, u^*(4)=5.$$

因此各城到 5 城的最短路线依次为

①→⑤,
②→③→④→⑤,
③→④→⑤,
④→⑤.

2.2 策略空间迭代法

2.2.1 迭代程序

(1) 先取一无回路的初始策略 $u_0(i)(i=1,2,\cdots,n)$,其中每个决策 $u_0(i)$ 代表由点 i 到达的下一个点.

(2) 假设已求得 $u_k(i)$,现由 $u_k(i)$ 按递推公式

$$\begin{cases} f_k(i) = C_{i,u_k(i)} + f_k(u_k(i)), & i=1,2,\cdots,n; \\ f_k(n+1) = 0 \end{cases}$$

求出 $f_k(i)$.

(3) 再由 $f_k(i)$ 按

$$\min_{1 \leqslant j \leqslant n+1} \{ C_{ij} + f_k(j) \}$$

求出 $u_{k+1}(i)$，即求使上式成立的相应决策作为 $u_{k+1}(i)$，但不能取含有 $C_{ii} = 0$ 的点作为 $u_{k+1}(i)$.

(4) 若经 $k+1$ 步迭代使 $u_{k+1}(i) = u_k(i)$ 对一切 $i = 1,2,\cdots,n$ 成立，则停止迭代，这时相应的 $\{u_k(i)\}$ 及 $f_k(i)$ 便是所求的最优策略及最优值．否则，返回步(2)继续迭代．

2.2.2 收敛性定理

定理 2 由迭代程序得到的函数列 $\{f_k(i)\}$，单调下降地一致收敛于函数方程(2-3)的解 $f(i)$.

定理 3 若初始策略 $\{u_0(i)\}$ 不构成回路，则每次迭代得到的策略 $\{u_k(i)\}$ 也都不构成回路．

例 2 利用策略空间迭代法重新求解例1中的问题．

解 (1) 取初始策略

$$u_0(1) = 5, \quad u_0(2) = 4, \quad u_0(3) = 5, \quad u_0(4) = 3.$$

(2) 由 $u_0(i)$ 求 $f_0(i)$:

$$f_0(i) = C_{i,u_0(i)} + f_0(u_0(i)), f_0(5) = 0,$$

即

$$f_0(1) = C_{15} + f_0(5) = C_{15} + 0 = 2,$$
$$f_0(3) = C_{35} + f_0(5) = C_{35} + 0 = 5,$$
$$f_0(4) = C_{43} + f_0(3) = 1 + 5 = 6,$$
$$f_0(2) = C_{24} + f_0(4) = 5 + 6 = 11.$$

(3) 再由 $f_0(i)$ 按

$$\min_{1 \leqslant j \leqslant 5} \{ C_{ij} + f_0(j) \}$$

求 $u_1(i)$. 令 $i = 1$ 得

$$\min\{C_{11} + f_0(1), C_{12} + f_0(2), C_{13} + f_0(3), C_{14} + f_0(4), C_{15} + f_0(5)\}$$
$$= \min\{0 + 2, 6 + 11, 5 + 5, 3 + 6, 2 + 0\} = 2.$$

由于最小值是在 $C_{15} + f_0(5)$ 处取得的，故得 $u_1(1) = 5$.

令 $i = 2,3,4$，类似可求得 $u_1(2) = 3, u_1(3) = 5, u_1(4) = 5$.

(4) 因为 $\{u_1(i)\} \neq \{u_0(i)\}$，故转(5)按以下步骤继续迭代．

(5) 由 $u_1(i)$ 按

$$f_1(i) = C_{i,u_1(i)} + f_1(u_1(i)), f_1(5) = 0$$

求 $f_1(i)$，令 $i = 1,2,3,4$ 可得

$$f_1(1) = 2, f_1(2) = 5.5, f_1(3) = 5, f_1(4) = 3.$$

(6) 再由 $f_1(i)$ 按

$$\min_{1 \leqslant j \leqslant 5} \{ C_{ij} + f_1(j) \}$$

求 $u_2(i)$，由此可分别求得

$$u_2(1) = 5, \quad u_2(2) = 3, \quad u_2(3) = 4, \quad u_2(5) = 5.$$

(7) 因为$\{u_2(i)\} \neq \{u_1(i)\}$,故转(8)按以下步骤继续迭代.

(8) 由$u_2(i)$可分别求得

$$f_2(1)=2, f_2(2)=4.5, f_2(3)=4, f_2(4)=3.$$

(9) 再由$f_2(i)$分别求得

$$u_2(1)=5, u_2(2)=3, u_2(3)=4, u_2(5)=5.$$

(10) 因为$\{u_2(i)\}=\{5,3,4,5\}=\{u_1(i)\}$,故停止迭代,最优策略为

$$\{u^*(i)\}=\{5,3,4,5\}.$$

由此,可得最优路线及相应的最短距离为

最优路线	相应的最短路线
①→⑤	2
②→③→④→⑤	4.5
③→④→⑤	4
④→⑤	3

所得结果与函数迭代法一致.

一般说来,策略迭代法的收敛速度要比函数迭代法快.

3 一些静态规划问题的动态规划解法

3.1 一类非线性规划问题的动态规划解法

考虑形如

$$\begin{cases}\text{opt}\{g_1(u_1) * g_2(u_2) * \cdots * g_n(u_n)\}; \\ \text{s.t.} \sum_{k=1}^{n} u_k = a, u_k \geqslant 0, k=1,2,\cdots,n\end{cases}$$

的一类非线性规划问题,其中opt取max或min,“$*$”表示一律取“+”或一律取“×”.

由于这类问题的特殊形式,它可以转化成下述的n阶段决策问题:

$$\begin{aligned}&\underset{u_1,u_2,\cdots,u_n}{\text{opt}}\{g_1(u_1) * g_2(u_2) * \cdots * g_n(u_n)\} \\ &= \underset{u_1}{\text{opt}}\{g_1(u_1) * \underset{u_2}{\text{opt}}[g_2(u_2) * \cdots * \underset{u_n}{\text{opt}}\, g_n(u_n)]\}.\end{aligned} \tag{3-1}$$

下面给出两种解法.

3.1.1 逆序解法

设状态变量为$x_1, x_2, \cdots, x_n$;决策变量为$u_1, u_2, \cdots, u_n$;它们之间的关系(即状态转移方程)为

$$u_n = x_n, u_{n-1} + x_n = x_{n-1}, \cdots, u_{k-1} + x_k = x_{k-1}, \cdots, u_1 + x_2 = x_1 = a. \tag{3-2}$$

令

$$\begin{cases} f_n(x_n) = \underset{u_n = x_n}{\text{opt}}\, g_n(u_n), \\ f_{n-k}(x_{n-k}) = \underset{0 \leqslant u_{n-k} \leqslant x_{n-k}}{\text{opt}} \{g_{n-k}(u_{n-k}) * f_{n-k+1}(x_{n-k+1})\} \\ \qquad = \underset{0 \leqslant u_{n-1} \leqslant x_{n-1}}{\text{opt}} \{g_{n-k}(u_{n-k}) * f_{n-k+1}(x_{n-k} - u_{n-k})\}, \\ \qquad k = 1,2,\cdots,n-1, \end{cases} \tag{3-3}$$

则由(3-1)式和(3-2)式知,原问题化为 n 个一元非线性规划问题(3-3).解(3-3)式便可得到原问题的解.

3.1.2　顺序解法

假设状态变量和决策变量符号不变,但状态转移方程改为

$$u_1 = x_1, u_2 + x_1 = x_2, \cdots, u_k + x_{k-1} = x_k, \cdots, u_n + x_{n-1} = x_n. \tag{3-4}$$

令

$$\begin{cases} f_1(x_1) = \underset{u_1 = x_1}{\text{opt}}\, g_1(u_1), \\ f_k(x_k) = \underset{0 \leqslant u_k \leqslant x_k}{\text{opt}} \{g_k(u_k) * f_{k-1}(x_{k-1})\} \\ \qquad = \underset{0 \leqslant u_k \leqslant x_k}{\text{opt}} \{g_k(u_k) * f_{k-1}(x_k - u_k)\}, \\ \qquad k = 2,3,\cdots,n. \end{cases} \tag{3-5}$$

利用(3-4)式解(3-5)式便可得到原问题的解.

例1　求解

$$\begin{cases} \max Z = u_1 \cdot u_2^2 \cdot u_3, \\ \text{s.t. } u_1 + u_2 + u_3 = 4, u_1, u_2, u_3 \geqslant 0. \end{cases}$$

解　用逆序解法.按(3-2)式和(3-3)式有

$$u_3 = x_3, u_2 + x_3 = x_2, u_1 + x_2 = x_1 = 4.$$

$$f_3(x_3) = \max_{u_3 = x_3} u_3 = x_3.$$

$$f_2(x_2) = \max_{0 \leqslant u_2 \leqslant x_2} \{u_2^2 \cdot f_3(x_2 - u_2)\} = \max_{0 \leqslant u_2 \leqslant x_2} \{u_2^2 \cdot (x_2 - u_2)\} = \max_{0 \leqslant u_2 \leqslant x_2} \{x_2 u_2^2 - u_2^3\},$$

由微分法易知,$u_2^* = \dfrac{2}{3}x_2, f_2(x_2) = \dfrac{4}{27}x_2^3$.

$$f_1(x_1) = \max_{0 \leqslant u_1 \leqslant 4} \{u_1 \cdot \frac{4}{27}(x_1 - u_1)^3\} = \max_{0 \leqslant u_1 \leqslant 4} \{u_1 \cdot \frac{4}{27}(4 - u_1)^3\},$$

又由微分法知,$u_1^* = 1, f_1(x_1) = 4$.于是有

$$u_1^* = 1, u_2^* = 2, u_3^* = 1; \max Z = 4.$$

例2　求解

$$\begin{cases} \max Z = 8x_1^2 + 4x_2^2 + x_3^3, \\ \text{s.t. } 2x_1 + x_2 + 10x_3 = b, x_1, x_2, x_3 \geqslant 0. \end{cases}$$

解　用顺序解法.设状态变量为 y_1, y_2, y_3;决策变量为 x_1, x_2, x_3.按(3-4)式和(3-5)式有

$$2x_1 = y_1, x_2 + y_1 = y_2, 10x_3 + y_2 = y_3 = b.$$

$$f_1(y_1) = \max_{0 \leqslant x_1 \leqslant y_1/2} 8x_1^2 = 8(y_{1/2})^2 = 2y_1^2, x_1^* = y_{1/2},$$

$$f_2(y_2) = \max_{0 \leqslant x_2 \leqslant y_2} \{4x_2^2 + f_1(y_2 - x_2)\} = \max_{0 \leqslant x_2 \leqslant y_2} \{4x_2^2 + 2(y_2 - x_2)^2\}.$$

由微分法易知 $x_2^* = y_2$.这时 $f_2(y_2) = 4y_2^2$.

$$f_3(y_3) = \max_{0 \leqslant x_3 \leqslant b/10} \{x_3^3 + f_2(b - 10x_3)\} = \max_{0 \leqslant x_3 \leqslant b/10} \{x_3^3 + 4(b - 10x_3)^2\},$$

易知 $x_3^3 + 4(b - 10x_3)^2$ 在$[0, b/10]$上是凸函数,故极大值只能在端点取得．注意,

$$\text{当 } x_3 = 0 \text{ 时}, \quad f_3 = 4b^2;$$

$$\text{当 } x_3 = b/10 \text{ 时}, \quad f_3 = b^3/10^3.$$

分别解 $4b^2 < b^3/10^3$ 和 $4b^2 > b^3/10^3$,得

$$\text{当 } b > 400 \text{ 时}, x_3^* = b/10, x_2^* = x_1^* = 0;$$

$$\text{当 } b < 400 \text{ 时}, x_1^* = 0, x_2^* = b, x_3^* = 0.$$

3.2　资源分配问题

资源分配问题是指:设有 m 种资源(如资金,原材料,设备,劳动力等),可以投入 n 种生产,试问将每种资源投入给各种生产各多少数量,才能使总的经济效益最大?下面仅讨论 $m = 1$ 和 $m = 2$ 两种情况,一般情形可类似讨论．

3.2.1　一种资源的分配问题

设有数量为 x_0 的某种资源,可以投入 n 种生产．若以数量为 x_i 的资源投入第 i 种生产所得的效益为 $g_i(x_i)$,试问如何分配这种资源,才能使总效益最大?

1. 数学模型

显然此问题的静态模型为

$$\max \sum_{i=1}^{n} g_i(x_i),$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{n} x_i = x_0, \\ x_i \geqslant 0, i = 1, 2, \cdots, n. \end{cases}$$

下面建立它的动态规划模型．为此,选取各动态参数如下:

阶段变量 k 选为 n 种生产的生产过程．

将投入第 k 种生产的资源数量 x_k 取作决策变量 u_k,即 $u_k = x_k$.

将投入第 k 种生产至第 n 种生产的总资源数量 x 取作状态变量．这时允许决策集合为

$$D_k(x) = \{u_k \mid 0 \leqslant u_k = x_k \leqslant x\}.$$

状态转移方程为

$$\tilde{x} = x - u_k = x - x_k,$$

其中 $\tilde{x}$ 表示投入第 $k+1$ 种生产至第 n 种生产的资源总量．

指标函数为

$$F_{k,n} = \sum_{i=k}^{n} g_i(x_i), k = 1,2,\cdots,n.$$

用 $f_k(x)$ 表示以数量 x 投入第 k 种生产至第 n 种生产时的最大效益，则由最优化原理，便可得到该问题的逆序动态规划方程为

$$\begin{cases} f_k(x) = \max\limits_{0 \leqslant u_k = x_k \leqslant x} \{g_k(x_k) + f_{k+1}(x - x_k)\}, \\ f_{n+1}(x) = 0, k = n, n-1, \cdots, 2, 1. \end{cases} \tag{3-6}$$

2. 数值解法

当 x 在 $[0, x_0]$ 上离散地变化时，可用逆序法求解(3-6)式；当 x 在 $[0, x_0]$ 上连续变化时，若 g_i 是凸函数，则(3-6)式的最优解必在 $[0, x_0]$ 的端点取得．这时(3-6)式仍可解，否则(3-6)式不能直接求解，但可采用离散化方法，给出一种求解方法．

用分点

$$x = 0, \Delta, 2\Delta, \cdots, m\Delta = x_0$$

将 $[0, x_0]$ 分成长为 Δ 的 m 个小区间，Δ 的大小可根据计算机的精度和容量来确定，并规定 x_k 和 $f_k(x)$ 都只在这些分点上取值．这样(3-6)式可写成

$$\begin{cases} f_k(x) = \max\limits_{0 \leqslant p \leqslant q} \{g_k(p\Delta) + f_{k+1}(x - p\Delta)\}, \\ f_{n+1}(x) = 0, x = q\Delta, q = 0, 1, \cdots, m. \end{cases} \tag{3-7}$$

由此可先求出 $f_n(x)$ 在各分点处的值，然后再由递推公式(3-7)求出 $f_{n-1}(x)$，$f_{n-2}(x), \cdots, f_1(x)$ 在各分点处的值．最后再反向追踪，求出最优决策，即从 $f_1(x_0)$ 求出最优决策 $u_1^* = x_1^*$，再从 $f_2(x - x_1^*)$ 求出 $u_2^* = x_2^*$，直到求出 $u_n^* = x_n^*$，则 $x_1^*, x_2^*, \cdots, x_n^*$ 就是最优分配方案，$f_1(x_0)$ 即为最大效益．

例3 某工业部门拟将五台设备分配给甲、乙、丙三个工厂，各厂得到该设备后，每年盈利情况如表3-1所示．

表3-1

设备台数	工厂		
	甲	乙	丙
0	0	0	0
1	3	5	4
2	7	10	6
3	9	11	11
4	12	11	12
5	13	11	12

试问分配给各工厂该种设备多少台，才能使各厂每年的总盈利最大？

解 这是一个典型的一维分配问题，其中 $x_0 = 5$；$g_1(x_1), g_2(x_2), g_3(x_3)$ 分别为甲，乙，丙工厂在得到 x_1, x_2, x_3 台设备后的盈利值；$x_k = 0,1,2,3,4,5(k = 1,2,3)$.

因为 $x = 0,1,2,3,4,5$，因此 x 在 $[0,5]$ 上离散变化，从而该问题的逆序动态规划方程(3-6) 变成

$$\begin{cases} f_k(x) = \max\limits_{0 \leq u_k = x_k \leq x} \{g_k(x_k) + f_{k+1}(x - x_k)\}, \\ f_4(x) = 0, k = 3,2,1. \end{cases}$$

(1) 当 $k = 3$ 时，则有

$$f_3(x) = \max_{0 \leq u_3 = x_3 \leq x} g_3(x_3).$$

由表 3-1 得

$$u_3 = 0, f_3(0) = 0; u_3 = 1, f_3(1) = 4;$$
$$u_3 = 2, f_3(2) = 6; u_3 = 3, f_3(3) = 11;$$
$$u_3 = 4, f_3(4) = 12; u_3 = 5, f_3(5) = 12.$$

所以 $u_3 = 4$ 或 5.

(2) 当 $k = 2$ 时，则有

$$f_2(x) = \max_{0 \leq u_2 = x_2 \leq x} \{g_2(x_2) + f_3(x - x_2)\}.$$

令 $x = 0,1,2,3,4,5$，得

$$f_2(0) = \max_{u_2 = x_2 = 0} \{g_2(0) + f_3(0)\} = 0, u_2(0) = 0.$$

$$\begin{aligned} f_2(1) &= \max_{0 \leq (u_2 = x_2) \leq 1} \{g_2(x_2) + f_3(1 - x_2)\} \\ &= \max\{g_2(0) + f_3(1), g_2(1) + f_3(0)\} \\ &= \max\{0 + 4, 5 + 0\} = 5, \end{aligned}$$

故 $u_2(1) = 1$.

$$\begin{aligned} f_2(2) &= \max\{g_2(0) + f_3(2), g_2(1) + f_3(1), g_2(2) + f_3(0)\} \\ &= \max\{0 + 6, 5 + 4, 10 + 0\} = 10, \end{aligned}$$

故 $u_2(2) = 2$.

同理可求出

$$f_2(3) = 14, \quad u_2(3) = 2;$$
$$f_2(4) = 16, \quad u_2(4) = 1 \text{ 或 } 2;$$
$$f_2(5) = 21, \quad u_2(5) = 2.$$

(3) 当 $k = 1$ 时，则有

$$f_1(x) = \max_{0 \leq (u_1 = x_1) \leq x} \{g_1(x_1) + f_2(x - x_1)\}.$$

注意，由 $f_1(x)$ 的定义知，$x = 5$. 于是有

$$\begin{aligned} f_1(5) = \max\{&g_1(0) + f_2(5), g_1(1) + f_2(4), g_1(2) + f_2(3), \\ &g_1(3) + f_2(2), g_1(4) + f_2(1), g_1(5) + f_2(0)\} \end{aligned}$$

$$= \max\{0+21, 3+16, 7+14, 9+10, 12+5, 13+0\} = 21,$$

故 $u_1(5) = 0$ 或 2.

最后,按步骤(3),(2),(1),并据 $f_1(5)$ 求出 $u_1^* = x_1^* = 0$ 或 2;再由 $f_2(x - x_1^*) = f_2(5)$ 或 $f_2(3)$ 得 $u_2^*(5) = 2$ 或 $u_2^*(3) = 2$;从而有 $u_3^* = 3$ 或 $u_3^* = 1$,即

$$u_1^* = 0, u_2^* = 2, u_3^* = 3;$$

或

$$u_1^* = 2, u_2^* = 2, u_3^* = 1.$$

亦即需分别给甲,乙,丙厂 0 台,2 台,3 台设备或 2 台,2 台,1 台设备才为最优方案. 这时总的最大盈利值为 21.

3.2.2　两种资源的分配问题

设有数量为 x_0 和 y_0 的两种资源,皆可投入 n 种生产. 若分别以数量为 x_i 和 y_i 的资源投入第 i 种生产所得效益为 $g_i(x_i, y_i)(i = 1,2,\cdots,n)$,试问如何分配这两种资源使之用于 n 种生产才能使总效益最大?

1. 数学模型

此问题的静态规划模型为

$$\max \sum_{i=1}^{n} g_i(x_i, y_i),$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{n} x_i = x_0, \sum_{i=1}^{n} y_i = y_0, \\ x_i, y_i \geqslant 0, i = 1,2,\cdots,n. \end{cases}$$

采用与一种资源类似的记号,可得逆序动态规划方程为

$$\begin{cases} f_k(x, y) = \max\limits_{\substack{0 \leqslant x_k \leqslant x \\ 0 \leqslant y_k \leqslant y}} \{g_k(x_k, y_k) + f_{k+1}(x - x_k, y - y_k)\}, \\ f_{n+1}(x, y) = 0, k = n, \cdots, 3,2,1. \end{cases} \tag{3-8}$$

2. 数值解法

当 x 和 y 分别在 $[0, x_0]$ 和 $[0, y_0]$ 上离散地变化时,可用逆序法求解(3-8)式. 否则将 $[0, x_0] \times [0, y_0]$ 用分点

$$x = 0, \Delta_1, 2\Delta_1, \cdots, m_1\Delta_1 = x_0,$$

$$y = 0, \Delta_2, 2\Delta_2, \cdots, m_2\Delta_2 = y_0$$

分成 $m_1 \times m_2$ 个小矩形,格点共有 $(m_1 + 1) \times (m_2 + 1)$ 个,并规定 (x_k, y_k) 和 $f_k(x_k, y_k)$ 只在这些格点上取值. 然后,利用(3-8)式计算 $f_k(x_k, y_k)$ 在这些格点上的值,再反向递推追踪,便可求出最优解

$$\{(x_1, y_1)^*, (x_2, y_2)^*, \cdots, (x_n, y_n)^*\}.$$

采用这种方法,当 n 或 m_1 和 m_2 很大时,会增加计算量和库存量. 为简化计算,可采用拉格朗日乘数法或逐次逼近法等降维方法.

3.3 复合系统工作的可靠性问题

设某种系统由 n 个部件"串联"而成,即一个部件失灵,整个系统就不能正常工作.为提高系统工作的可靠性,在每个部件上都装有主要元件的备用件及自动投入装置.备用件越多,系统正常工作的可靠性就越大.但备用件过多时,系统的成本、质量和体积等均相应增大,工作精度也会降低.因此,这里的优化问题就变成了如何配备各部件的备用件,使系统工作的可靠性最大的问题.

设第 $i(i=1,2,\cdots,n)$ 个部件上装有 z_i 个备用件,那么 z_i 的概率 $P_i(z_i)$,恰好表示了第 i 个部件正常工作的程度.于是系统正常工作的可靠程度便可用

$$P(z_1,z_2,\cdots,z_n)=\prod_{i=1}^{n}P_i(z_i)$$

来度量.

3.3.1 静态模型

设第 i 个部件上装有 z_i 个备用件,每个备用件的费用为 C_i,质量为 W_i,且总费用不超过 C,总质量不超过 W,则系统正常工作的静态规划模型为

$$\max P(z_1,z_2,\cdots,z_n)=\prod_{i=1}^{n}P_i(z_i),$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}C_iz_i\leqslant C,\ \sum_{i=1}^{n}W_iz_i\leqslant W,\\ z_i\geqslant 0\ \text{为正整数},\ i=1,2,\cdots,n.\end{cases}$$

这是一个非线性整数规划问题,一般求解较困难.

3.3.2 动态模型

将 n 个部件看作 n 个阶段,用 k 表示第 k 个阶段.

由于问题有两个约束条件,故选二维状态变量:

x_k 表示由第 k 个部件到第 n 个部件使用的总费用;

y_k 表示由第 k 个部件到第 n 个部件具有的总质量.

决策变量 u_k 取第 k 个部件上装配的备用件数,即 $u_k=z_k$.

决策集合,状态转移方程和指标函数依次为

$$D_k(x_k,y_k)=\left\{u_k\mid 0\leqslant u_k\leqslant\min\left(\frac{x_k}{C_k},\frac{y_k}{W_k}\right)\right\},$$

(且 u_k 取非负整数),

$$\begin{cases}x_{k+1}=x_k-u_kC_k,\\ y_{k+1}=y_k-u_kW_k\end{cases}$$

和

$$F_{k,n}=\prod_{i=k}^{n}P_i(z_i),\ k=1,2,\cdots,n.$$

如用 $f_k(x_k, y_k)$ 表示由第 k 个到第 n 个部件正常工作的最大可靠程度，则由最优化原理，便得逆序动态规划方程为

$$f_k(x_k, y_k) = \max_{u_k \in D_k} \{P_k(u_k) \cdot f_{k+1}(x_k - u_k C_k, y_k - u_k W_k)\},$$
$$f_{n+1}(x_{n+1}, y_{n+1}) = 1, k = n, \cdots, 3, 2, 1.$$

3.4 背包问题

3.4.1 背包问题及其数学模型

所谓背包问题是指这样的问题：一个人带一个背包上山，背包可携带物品的总质量为 W. 现有 n 种物品，第 i 种物品每件的质量为 W_i，若带 x_i 件第 i 种物品，则可获价值为 $C_i(x_i)$，试问带哪几种物品各多少件，才能使获得的总价值最大？

背包问题实际上是一类货物装运问题．它的静态规划模型显然为

$$\begin{cases} \max \sum_{i=1}^{n} C_i(x_i), \\ \text{s.t.} \sum_{i=1}^{n} W_i x_i \leqslant W, \quad x_i \text{ 为非负整数}, i = 1, 2, \cdots, n. \end{cases}$$

一般说来，这是一个非线性整数规划问题，求解比较困难．下面建立它的动态规划模型．

设状态变量 x 表示可装载的第1种至第 k 种物品的质量(也叫装载能力)；决策变量 u_k 表示第 k 种物品的装载数量(件数) x_k，即 $u_k = x_k$. 于是，决策集合和状态转移方程分别为

$$D_k(x) = \{u_k \mid 0 \leqslant u_k = x_k \leqslant [x/W_k]\},$$

$$\tilde{x} = x - W_k x_k,$$

其中 $[x/W_k]$ 表示对 x/W_k 取整，$\tilde{x}$ 表示可装载的第1种至第 $k-1$ 种物品的质量．指标函数为

$$F_{1,k} = \sum_{i=1}^{k} C_i(x_i).$$

用 $f_k(x)$ 表示装载能力为 x 时，可装载的第1种至第 k 种物品所获得的最大价值，则由最优化原理便得顺序动态规划方程为

$$\begin{cases} f_k(x) = \max\limits_{x_k = 0, 1, \cdots, [x/W_k]} \{C_k(x_k) + f_{k-1}(x - W_k x_k)\}, \\ f_0(x) = 0, k = 1, 2, \cdots, n. \end{cases} \tag{3-9}$$

仿此可知，逆序动态规划方程为

$$\begin{cases} f_k(x) = \max\limits_{x_k = 0, 1, \cdots, [x/W_k]} \{C_k(x_k) + f_{k+1}(x - W_k x_k)\}, \\ f_{n+1}(x) = 0, k = n, \cdots, 2, 1, \end{cases}$$

其中 x 表示可装载的第 k 种至第 n 种物品的质量，$f_k(x)$ 表示装载能力为 x 时，可装载的第 k 种至第 n 种物品所获得的最大价值．

例 4　假设背包的装载能力为100，现有 4 种物品，每种物品每件的质量和价值如表 3-2 所示，试求最优装载方案和最大价值．

表 3-2

物品种类	质　　量	价　　值
1	40	40
2	30	25
3	15	10
4	8	5

解　用 v_i 表示第 i 种物品每件的价值，则(3-9) 式变成

$$\begin{cases} f_k(x) = \max\limits_{x_k=0,1,\cdots,[x/W_k]} \{v_k x_k + f_{k-1}(x - W_k x_k)\}, \\ f_0(x) = 0,\ k = 1,2,3,4. \end{cases} \tag{3-10}$$

注意，x 在$[0,100]$上可连续取值，为了能用(3-10) 式进行计算，将 x 离散化，取 $x = 100,95,90,\cdots,10,5,0$，如表 3-3 中的第 1 列所示．

按顺序解法求解．先计算 $f_1(100)$．由于$[x/W_1] = [100/40] = 2$，故 $x_1 = 0,1,2$，从而

$$f_1(100) = \max\{40 \times 0 + 0, 40 \times 1 + 0, 40 \times 2 + 0\} = 80.$$

仿此可求出 $f_1(95), f_1(90), \cdots, f_1(5), f_1(0)$，列入表 3-3 的第 2 列中．

再计算 $f_2(100)$．尽管$[x/W_2] = [100/30] = 3$，但不取 $x_2 = 3$，因为三件第 2 种物品的质量为 90，而价值只有 75，不如取两件第 1 种物品，价值可达 80．因此，$x_2 = 0,1,2$，从而

$$\begin{aligned} f_2(100) &= \max\{0 + f_1(100), 25 + f_1(70), 50 + f_1(40)\} \\ &= \max\{0 + 80, 25 + 40, 50 + 40\} = 90. \end{aligned}$$

同理，可求出 $f_2(95), f_2(90), \cdots, f_2(5), f_2(0)$，列入表 3-3 的第 3 列中．

如同分析 $f_2(100)$ 那样，对 $f_3(100)$ 只须取 $x_3 = 0,1$，故

$$\begin{aligned} f_3(100) &= \max\{0 + f_2(100), 10 + f_2(85)\} \\ &= \max\{0 + 90, 10 + 80\} = 90. \end{aligned}$$

类似地，可求出 $f_3(95), f_3(90), \cdots, f_3(5), f_3(0)$，列入表 3-3 的第 4 列中．

最后，计算 $f_4(100)$，这时 $x_4 = 0,1,2,3$，故

$$\begin{aligned} f_4(100) &= \max\{0 + f_3(100), 5 + f_3(92), 10 + f_3(84), 15 + f_3(76)\} \\ &= \max\{0 + 90, 5 + 80, 10 + 80, 15 + 65\} = 90. \end{aligned}$$

仿此可求出 $f_4(95), f_4(90), \cdots, f_4(5), f_4(0)$，列入表 3-3 的第 5 列中．

按定义 $f_4(100) = 90$ 为最大价值．下面用反向追踪的办法求最优装载方案，即最优策略．

表 3-3

x	$f_1(x)$			$f_2(x)$			$f_3(x)$		$f_4(x)$			
	$x_1=0$	1	2	$x_2=0$	1	2	$x_3=0$	1	$x_4=0$	1	2	3
100			80			90	90	90	90		90	
95			80	80				90	90			
90			80	80			80			85		
85			80	80			80		80			
80			80	80			80		80			
75		40			65		65		65			
70		40			65		65		65			
65		40				50	50	50		55		55
60		40				50	50	50	50		50	
55		40		40				50	50			
50		40		40			40			45		
45		40		40			40		40			
40		40		40			40		40			
35	0				25		25		25			
30	0				25		25		25			
25	0			0				10		15		15
20	0			0				10	10		10	
15	0			0				10	10			
10	0			0				0		5		
5	0			0			0		0			
0	0			0			0		0			

为保证最大价值,先确定 x_4:

$$x_4(100)=0,\quad 剩余容量为 100;$$

$$x_4(100)=2,\quad 剩余容量为 84.$$

按剩余容量确定 x_3:

$$x_3(100)=0,\quad 剩余容量为 100;$$

$$x_3(100)=1,\quad 剩余容量为 85;$$

$$x_3(84)=0,\quad 剩余容量为 84.$$

这里84可近似看做85,以便利用表3-3(以下类似情况不再声明).再按剩余容量确定 x_2:

$$x_2(100)=2,\quad 剩余容量为 40;$$

$$x_2(85)=0,\quad 剩余容量为 85;$$

$$x_2(84) = 0, \quad \text{剩余容量为 } 84.$$

最后按剩余容量确定 x_1:

$$x_1(40) = 1, \quad \text{剩余容量为 } 0;$$
$$x_1(85) = 2, \quad \text{剩余容量为 } 5;$$
$$x_1(84) = 2, \quad \text{剩余容量为 } 4.$$

由此,得三种最优装载方案

$$\{x_1 = 2, x_2 = 0, x_3 = 0, x_4 = 2\},$$
$$\{x_1 = 2, x_2 = 0, x_3 = 1, x_4 = 0\},$$
$$\{x_1 = 1, x_2 = 2, x_3 = 0, x_4 = 0\}.$$

三种最优装载方案的最大价值都是 90.

3.4.2　价值函数 $C_k(x_k)$ 为线性函数的情形

由例 4 可知,利用顺序动态规划方程(3-9)式求解背包问题. 当 W 较大时,存储量是很大的. 当价值函数 $C_k(x_k)$ 为线性函数 $C_k(x_k) = v_k x_k (k = 1,2,\cdots,n)$ 时(例 4 就是这种情形),其中 v_k 表示第 k 种物品每件的价值,可以建立一种比(3-9)式简单的模型.

不妨假设诸质量 W_i 都是正整数,且所有 W_i 的最大公约数为 1,只要重新调整质量单位,这两个要求就都能达到.

将每次装载一件物品作为一个阶段,便形成一个不定期的多阶段决策问题.

取状态变量为每个阶段可供装载的物品的质量. 决策变量取决定装载物品的种类. 若决定装第 i 种物品,则剩下可供装载的物品质量为 $\tilde{W} = W - W_i$,这就是状态转移方程. 状态决策集合为 $D = \{i \mid i = 1,2,\cdots,n\}$. 得到的价值为 v_i.

用 $f(W)$ 表示装载物品质量为 W 时所能获得的最大价值,则由最优化原理便得

$$\begin{cases} f(W) = \max\limits_{i=1,2,\cdots,n} \{v_i + f(W - W_i)\}; \\ f(W) = 0, 0 \leqslant W \leqslant \underline{W} - 1, \end{cases} \tag{3-11}$$

其中 $\underline{W} = \min\limits_{i=1,2,\cdots,n} \{W_i\}$,并规定 $f(-W) = -\infty$,这就是问题的动态规划模型.

(3-11)式比(3-9)式简单,因为每次只须递推地计算一个函数最优值.

例 5　计算表3-4所给数据的装载问题,其中 $W = 22$.

表 3-4

货物种类 i	质量 W_i	价值 v_i	v_i/W_i
1	3	7	7/3
2	6	16	8/3
3	7	19	19/7
4	5	15	3

解　用 $p(W)$ 表示可供装载货物的质量为 W 时,使(3-11)式成立的第 i 种货

物.

因为$\underline{W} = \min\{3,6,7,5\} = 3$,故由(3-11)式中的边界条件知

$$f(0) = f(1) = f(2) = 0.$$

再由递推公式(3-11)依次计算$f(3),f(4)$,直至$f(22)$.

$$\begin{aligned} f(3) &= \max\{v_1 + f(0), v_2 + f(-3), v_3 + f(-4), v_4 + f(-2)\} \\ &= \max\{7 + 0, 16 - \infty, 19 - \infty, 15 - \infty\} = 7,\ p(3) = 1, \\ f(4) &= \max\{v_1 + f(1), v_2 + f(-2), v_3 + f(-3), v_4 + f(-1)\} \\ &= \max\{7 + 0, 16 - \infty, 19 - \infty, 15 - \infty\} = 7,\ p(4) = 1, \\ f(5) &= \max\{7 + f(2), 16 + f(-1), 19 + f(-2), 15 + f(0)\} \\ &= \max\{7 + 0, 16 - \infty, 19 - \infty, 15 + 0\} = 15,\ p(5) = 4. \end{aligned}$$

类似地可算出

$f(6) = 16, p(6) = 2.$　　$f(7) = 19, p(7) = 3.$
$f(8) = 22, p(8) = 1$ 或 4.　　$f(9) = 23, p(9) = 1$ 或 2.
$f(10) = 30, p(10) = 4.$　　$f(11) = 31, p(11) = 2$ 或 4.
$f(12) = 34, p(12) = 3$ 或 4.　　$f(13) = 37, p(13) = 1$ 或 4.
$f(14) = 38, p(14) = 1$ 或 2,3,4.　　$f(15) = 45, p(15) = 4.$
$f(16) = 46, p(16) = 2$ 或 4.　　$f(17) = 49, p(17) = 3$ 或 4.
$f(18) = 52, p(18) = 1$ 或 4.　　$f(19) = 53, p(19) = 1$ 或 2,3,4.
$f(20) = 60, p(20) = 4.$　　$f(21) = 61, p(21) = 2$ 或 4.
$f(22) = 64, p(22) = 3$ 或 4.

由f的定义知,$f(22) = 64$为最大价值.$p(22) = 3$或4表示可供装载的货物质量为22,需装第3和第4种货物,而这两种货物每件的质量分别为7和5(见表3-4).因此,可装第3种货物1件,第4种货物3件,而第1,2种货物都不装,即最优装载方案为

$$x_1 = x_2 = 0,\quad x_3 = 1,\quad x_4 = 3.$$

4　确定型动态规划应用举例

4.1　另一类资源分配问题

3.1节中研究的资源分配问题是一次性生产问题,即不考虑收回剩余资源进行再生产.现在研究另一类资源的分配问题,即每次生产后,需将剩余资源回收后进行再生产.

为简单起见,只考虑一种资源、两种生产的情况.

设有数量为x_0的某种资源,可以投入A,B两种生产.假设投入A,B两种生产时,所得效益S_1,S_2与投入资源数量y,z的关系分别为$S_1 = g(y)$和$S_2 = h(z)$.每次生产后,可将剩余资源回收进行再生产,设回收率分别为a和$b(0 < a,b < 1)$.

现在制定一个 n 年生产计划，其目的是将每次回收的资源重新分配给两种生产，使总效益最大．

4.1.1　数学模型

这是一个 n 阶段决策问题．阶段变量 k 取作年度，即 $k=1,2,\cdots,n$.

状态变量 x_k 取作第 k 年年初拥有的资源数量．

决策变量 u_k 取作第 k 年年初分配给 A 种生产的资源数量，这时分配给 B 种生产的资源数量自然为 x_k-u_k.

允许决策集合及状态转移方程分别为

$$D_k=\{u_k\mid 0\leqslant u_k\leqslant x_k\},$$
$$x_{k+1}=au_k+b(x_k-u_k).$$

阶段效益即为第 k 年度的效益，故为

$$d(x_k,u_k)=g(u_k)+h(x_k-u_k).$$

指标函数显然为

$$F_{k,n}=\sum_{j=k}^{n}[g(u_j)+h(x_j-u_j)],\quad k=1,2,\cdots,n.$$

用 $f_k(x_k)$ 表示从状态 x_k 出发，采用最优策略，到第 n 年生产结束时的最大效益由最优化原理，得逆序动态规划方程为

$$\begin{cases}f_k(x_k)=\max\limits_{0\leqslant u_k\leqslant x_k}\{g(u_k)+h(x_k-u_k)+f_{k+1}(au_k+b(x_k-u_k))\}\\ f_{n+1}(x_{n+1})=0,k=n,\cdots,2,1.\end{cases}\tag{4-1}$$

由于 x_k 和 u_k 可以连续变化，故用 x 和 y 分别表示之．上述递推公式可写成

$$\begin{cases}f_k(x)=\max\limits_{0\leqslant y\leqslant x}\{g(y)+h(x-y)+f_{k+1}(ay+b(x-y))\},\\ f_{n+1}(x)=0,k=n,\cdots,2,1.\end{cases}\tag{4-2}$$

4.1.2　模型(4-1)的解法

一般说来，模型(4-2)的每一次迭代都是一个非线性规划，求解较为复杂．但当 $g(y)$ 和 $h(y)$ 都是凸函数时，求解是容易的．

定理 1　若 $g(y)$ 和 $h(y)$ 都是凸函数，且 $g(0)=h(0)=0$，则动态规划方程(4-2)的最优策略 y，在每个阶段总取区间 $[0,x]$ 的端点．

例 1(机器负荷问题)　设有 1 000 台机器，可以在高低两种不同负荷下进行生产．假设机器在高负荷下生产时，产品的年产量 S_1 与投入的机器数量 y 的关系为 $S=8y$，机器的完好率为 0.7；机器在低负荷下生产时，产品的年产量 S_2 与投入的机器数量 z 的关系为 $S_2=5z$，机器的完好率为 0.9，要求制定一个 5 年的生产计划．问每年开始时如何重新分配完好的机器在两种负荷下工作的数量，才能使 5 年内总产量最高．

解　显然，这个问题是上述资源分配问题的特例，其中 $x_0=1\,000$, $n=5$, $g=8y$, $h=5z$, $a=0.7$, $b=0.9$，这时(4-1)式变成

$$\begin{cases} f_k(x_k) = \max\limits_{0 \leq u_k \leq x_k} \{8u_k + 5(x_k - u_k) + f_{k+1}(0.7u_k + 0.9(x_k - u_k))\}, \\ f_6(x_6) = 0, k = 5,4,3,2,1. \end{cases}$$

由于 $g = 8y, h = 5z$ 都是线性函数，当然是凸函数，故可利用定理1来解．

当 $k = 5$ 时，

$$f_5(x_5) = \max_{0 \leq u_5 \leq x_5} \{8u_5 + 5(x_5 - u_5)\} = \max_{0 \leq u_5 \leq x_5} \{3u_5 + 5x_5\}.$$

因为 $3u_5 + 5x_5$ 是 u_5 的线性单调增函数，故当 $u_5 = x_5$ 时取最大值(以下类似情况不再说明)，所以

$$f_5(x_5) = 8x_5.$$

当 $k = 4$ 时，

$$f_4(x_4) = \max_{0 \leq u_4 \leq x_4} \{8u_4 + 5(x_4 - u_4) + 8[0.7u_4 + 0.9(x_4 - u_4)]\}$$
$$= \max_{0 \leq u_4 \leq x_4} \{1.4u_4 + 12.2x_4\} = 13.6x_4, u_4 = x_4;$$

当 $k = 3$ 时，

$$f_3(x_3) = \max_{0 \leq u_3 \leq x_3} \{8u_3 + 5(x_3 - u_3) + 13.6[0.7u_3 + 0.9(x_3 - u_3)]\}$$
$$= \max_{0 \leq u_3 \leq x_3} \{17.52u_3 + 17.24(x_3 - u_3)\} = 17.52x_3, u_3 = x_3;$$

当 $k = 2$ 时，

$$f_2(x_2) = \max_{0 \leq u_2 \leq x_2} \{8u_2 + 5(x_2 - u_2) + 17.52[0.7u_2 + 0.9(x_2 - u_2)]\}$$
$$= \max_{0 \leq u_2 \leq x_2} \{20.26u_2 + 20.768(x_2 - u_2)\} = 20.768x_2, u_2 = 0;$$

当 $k = 1$ 时，

$$f_1(x_1) = \max_{0 \leq u_1 \leq x_1} \{8u_1 + 5(x_1 - u_1) + 20.768[0.7u_1 + 0.9(x_1 - u_1)]\}$$
$$= \max_{0 \leq u_1 \leq x_1} \{22.5376u_1 + 23.6912(x_1 - u_1)\} = 23.7x_1, u_1 = 0.$$

由此可知，最优策略为

$$\{u_1^* = 0, u_2^* = 0, u_3^* = x_3, u_4^* = x_4, u_5^* = x_5\},$$

即头两年把年初的完好机器全部投入低负荷生产，后三年则全部投入高负荷生产，这样使得最高产量

$$f_1(x_1) = 23.7x_1 = 23700 \text{ 台}.$$

利用状态方程可以求出每年年初尚有的完好机器数量分别为

$x_1 = 1000$ 台，$x_2 = 0.7u_1 + 0.9(x_1 - u_1) = 900$ 台，

$x_3 = 0.7u_2 + 0.9(x_2 - u_2) = 810$ 台，$x_4 = 0.7u_3 + 0.9(x_3 - u_3) = 570$ 台，

$x_5 = 0.7u_4 + 0.9(x_4 - u_4) = 397$ 台，$x_6 = 0.7u_5 + 0.9(x_5 - u_5) = 278$ 台．

上面的问题始端状态 $x_1 = 1000$ 台是给定的，但终端状态 x_6 没有限制，这样对生产显然不利．因此，通常对终端是有限制的，例如规定 $x_6 = 500$ 台，即5年后尚须保存完好机器500台，这时如何安排生产，才能使总产量最高？这个问题是不难解决的．

4.2 连轧机操作问题

所谓连轧机操作问题是指这样的问题:假设轧机有 n 个彼此独立的轧辊,钢坯通过每个轧辊轧制后,厚度都要变小,即都有一定的压下率,最后轧制成所要求的厚度. 在轧制过程中,通过每一道轧辊的压下率是可以选择的,但要受工艺条件的限制,压下率不能太大,否则会使作用于轧辊的反抗力矩过大而损坏机器,同时还会使轧制时间过长而影响以后各轧辊的轧制. 因此,各个轧辊的压下率就构成不同的轧制策略,不同的轧制策略对应着不同的轧制结果,因而也就对应着不同的产量和质量. 那么,如何求出最优的压下率,使得在保证产品质量的前提下,钢材通过的终速度为最大.

4.2.1 数学模型

将 n 个轧辊的轧制次序 $k = 1,2,\cdots,n$ 取作阶段变量.

用 y_k, x_k 分别表示轧材通过第 k 个轧辊的入口和出口厚度. 并取 x_k 作为第 k 段的状态变量,而将决策变量 u_k 取作 y_k,即 $u_k = y_k$. 于是,决策集合为

$$D_k = \{u_k \mid x_k \leqslant u_k = y_k \leqslant p\},$$

其中 p 为第一个轧辊的入口厚度,并注意 $y_k = x_{k-1}$,故状态转移方程为

$$x_{k-1} = u_k = y_k.$$

用 $t(x_k, y_k)$ 表示当轧材的入口、出口厚度分别为 y_k, x_k 时,轧材通过第 k 个轧辊的最大速度. $T_k(x_k)$ 表示轧材通过第 k 个轧辊的出口厚度为 x_k 时,轧材通过前 k 个轧辊的最大速度.

于是,由最优化原理,并注意轧材通过轧制系统任何两部分的速度不能超过这两部分速度的最小者,便得该问题的顺序动态规划方程为

$$\begin{cases} T_k(x_k) = \max\limits_{x_k \leqslant y_k \leqslant p} \{\min[T_{k-1}(y_k), t(x_k, y_k)]\}, \\ T_1(x_1) = t(x_1, p), k = 2,3,\cdots,n. \end{cases} \tag{4-3}$$

例 2 表4-1给出了对应不同的 x_k 和 y_k 的 $t(x_k, y_k)$ 的值, $p = 10$, 试计算相应的 $T_k(x_k), k = 1,2,3,4,5$.

解 由边界条件知 $T_1(x_1) = t(x_1, 10)$, 而 $x_1 = 10,9,\cdots,1$, 于是 $T_1(x_1, 10)$ 的值如同表 4-1 中的第 2 行,把它记入表 4-3 中的第 2 行,这时 $u_1(x_1)$ 无意义.

再由递推公式(4-3) 计算 $T_2(x_2)$,这时

$$T_2(x_2) = \max_{x_2 \leqslant y_2 \leqslant 10} \{\min[T_1(y_2), t(x_2, y_2)]\}.$$

于是,由表 4-1 知,

$$\begin{aligned} T_2(10) &= \max_{10 \leqslant y_2 \leqslant 10} \{\min[T_1(y_2), t(10, y_2)]\} \\ &= \min[T_1(10), t(10,10)] = \min[1.00, 1.00] \\ &= 1.00. \end{aligned}$$

$$T_2(9) = \max_{9 \leqslant y_2 \leqslant 10} \{\min[T_1(y_2), t(9, y_2)]\}$$

表4-1　$t(x_k,y_k)$

y_k	x_k									
	10	9	8	7	6	5	4	3	2	1
10	1.00	0.95	0.89	0.84	0.77	0.71	0.63	0.55	0.45	0.32
9		1.00	0.95	0.88	0.82	0.74	0.67	0.57	0.47	0.33
8			1.00	0.94	0.87	0.79	0.71	0.61	0.50	0.35
7				1.00	0.93	0.84	0.76	0.66	0.55	0.37
6					1.00	0.91	0.82	0.71	0.58	0.41
5						1.00	0.89	0.77	0.63	0.45
4							1.00	0.87	0.71	0.50
3								1.00	0.82	0.57
2									1.00	0.71
1										1.00

$$= \max\begin{Bmatrix}\min[T_1(9),t(9,9)]\\ \min[T_1(10),t(9,10)]\end{Bmatrix} = \max\begin{Bmatrix}\min[0.95,1.00]\\ \min[1.00,0.95]\end{Bmatrix}$$
$$= 0.95,u_2(9) = 9 \text{或} 10.$$

类似地,可求出 $T_2(8),\cdots,T_2(1)$.

为了简化上述计算过程,可采用列表的办法直观地进行.表4-2给出了 $T_2(3)$ 的计算结果.一般规则是:为计算 $T_k(x_k)$,取表4-1中右边第 k 列的元素与 $T_{k-1}(x_k)$ 中相应的元素成对地进行比较后取最小值,最后取这些最小值的最大者就是 $T_k(x_k)$.由表4-2知,$T_2(3) = 0.71,u_2(3) = 6$ 或5.亦即为使轧材通过第2个轧辊的出口厚度为3,必须使轧辊的入口厚度为6或5,才能使轧材的出口速度达到最大值0.71.

表4-3记录了 $x_k = 10,9,8,\cdots,1,k = 1,2,3,4,5$ 的 $T_k(x_k)$ 和 u_k 的各种结果.

表4-2

y	表4-1中第3列	T_1 的行	最小值	解
10	0.55	1.00	0.55	
9	0.57	0.95	0.57	
8	0.61	0.89	0.61	
7	0.66	0.84	0.66	
6	0.71	0.77	0.71	6
5	0.77	0.71	0.71	5
4	0.87	0.63	0.63	
3	1.00	0.55	0.55	

表 4-3

x	10	9	8	7	6	5	4	3	2	1
$T_1(x)$	1.00	0.95	0.89	0.84	0.71	0.71	0.63	0.55	0.45	0.32
$u_1(x)$	-	-	-	-	-	-	-	-	-	-
$T_2(x)$	1.00	0.95	0.95	0.89	0.87	0.84	0.77	0.71	0.63	0.55
$u_2(x)$	10	10,9	9	8	8	7	6	5,6	4,5	3
$T_3(x)$	1.00	0.95	0.95	0.94	0.89	0.87	0.84	0.77	0.71	0.63
$u_3(x)$	10	10,9	8,9	8	7	6	5	5,4	3,4	2
$T_4(x)$	1.00	0.95	0.95	0.94	0.97	0.89	0.87	0.84	0.77	0.71
$u_4(x)$	10	10,9	8,9	7,8	7	6	5	4	3	2
$T_5(x)$	1.00	0.95	0.95	0.94	0.93	0.91	0.89	0.87	0.82	0.71
$u_5(x)$	10	10,9	8,9	7,8	7,6	6	5	4	3	2,1

4.2.2 递推公式(4-3)的图像解法

由于入口和出口厚度都是连续变化的，故可将 $t(x_k, y_k)$ 写成 $t(x, y)$，并将 $z = t(x, y)$ 看做以 x 为参数的曲线簇，它对 y 是单调减少的，对 x 是单调增加的．又曲线 $z = T_{k-1}(y)$ 是单调增加的．在实际问题中，这些要求都是满足的．

当 $t(x, y)$ 及 $T_{k-1}(y)$ 为已知时，求出曲线 $z = t(x, y)$ 与 $z = T_{k-1}(y)$ 的交点的 y 坐标 $y = u_k(x)$，如图 4-1 所示，便得到

$$T_k(x) = t(u_k(x), x).$$

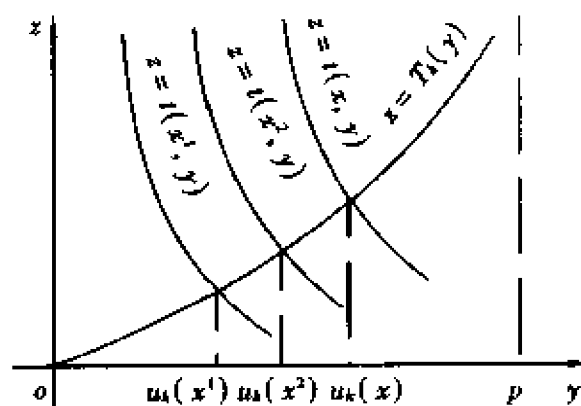


图 4-1

4.3 设备更新问题

在工农业生产和国防建设中，经常遇到因设备陈旧或损坏需要更换新的问题．

那么,一台设备使用多少年后更新,才能使总的效益最大?这就是设备更新问题.

4.3.1 几个基本概念

用 $r(t)$ 表示役令(已使用过的年限)为 t 年的一台设备再继续使用一年可得到的经济收入,称 $r(t)$ 为**效益函数**.通常 $r(t)$ 是减函数.

用 $u(t)$ 表示役令为 t 的一台设备再继续使用一年的维修费用,称 $u(t)$ 为**维修费用函数**.通常 $u(t)$ 是增函数.

用 $C(t)$ 表示卖掉一台役令为 t 的旧设备,买进一台新设备(役令为0年)的纯支出费用,称 $C(t)$ 为**更新费用函数**.$C(t)$ 总是非负的,并在价格不变的条件下是增函数.

4.3.2 数学模型

假设某工厂需连续营运 n 年,该厂有一台某种设备的 $r(t),u(t),C(t)$ 均已知.选取动态参数如下:

阶段变量 $k(k=1,2,\cdots,n)$ 选为营运年数.

状态变量 x_k 表示第 k 年年初设备的役令.

决策变量 u_k 表示第 k 年年初是继续使用旧设备(用 K 表示),还是使用更新后的新设备(用 P 表示),即 $u_k=K$ 或 P.

状态转移方程为

$$x_{k+1}=\begin{cases}x_k+1, & u_k=K;\\ 1, & u_k=P.\end{cases}$$

阶段指标函数

$$d(x_k,u_k)=\begin{cases}r(x_k)-u(x_k), & u_k=K;\\ r(0)-u(0)-C(x_k), & u_k=P.\end{cases}$$

指标函数为

$$F_{k,n}=\sum_{j=k}^{n}d(x_j,u_j),\quad k=1,2,\cdots,n.$$

用 $f_k(x_k)$ 表示从第 k 年年初开始,使用一台役令为 x_k 的设备,到第 n 年年末的最大纯收益.由最优化原理,得逆序动态规划方程

$$\begin{cases}f_k(x_k)=\max\limits_{u_k=K或P}\{d(x_k,u_k)+f_{k+1}(x_{k+1})\}\\ \qquad =\max\begin{cases}r(x_k)-u(x_k)+f_{k+1}(x_k+1), & u_k=K;\\ r(0)-u(0)-C(x_k)+f_{k+1}(1), & u_k=P,\end{cases}\\ f_{n+1}(x_{n+1})=0,k=n,\cdots,2,1.\end{cases}\tag{4-4}$$

如果开始时役令为0,则 x_1 的可达状态为0,$f_1(0)$ 为所求最大效益;如果开始时役令为 m,则 x_1 的可达状态为 m,$f_1(m)$ 为所求最大效益.最优策略可按 u_1 是取 K 还是取 P 才产生 $f_1(0)$ 定出 u_1^*,再由产生 $f_1(0)$ 的表达式确定 $f_2(x_2)$,然后再按 u_2 是取 K 还是取 P 才产生 $f_2(x_2)$ 定出 u_2^*.依此继续下去,便可求出最优更新策略

$\{u_1^*, u_2^*, \cdots, u_n^*\}$.

例 3 某工厂的某型号机床的年均维修费用及效益指标如表4-4所示(单位为千元).

表4-4

项　　目	役令/年					
	0	1	2	3	4	5
$v(t)$/千元	5	4.5	4	3.75	3	2.5
$u(t)$/千元	0.5	1	1.5	2	2.5	3

购买一台同型号的新机床价格为5 000元. 如厂方将该机床出售,其价格如表4-5所示.

表4-5

役令/年	0	1	2	3	4	5
价格/千元	4.5	4	3.5	3	2.5	2

该厂在1996年有一台新机床,试给出至2000年年底该机床的最优更新策略及最优效益值.

解 这个问题正是递推公式(4-4)所描述的类型,其中 $n = 5$, $r(t)$ 及 $u(t)$ 分别如表4-4及4-5所示. $C(t)$ 如表4-6所示.

表4-6

役令/年	0	1	2	3	4	5
$C(t)$/千元	0.5	1	1.5	2	2.5	3

当 $k = 5$ 时,递推公式(4-4)变成

$$f_5(x_5) = \max\begin{cases} r(x_5) - u(x_5), & u_5 = K; \\ r(0) - u(0) - C(x_5), & u_5 = P. \end{cases}$$

这时 x_5 的可达状态为1,2,3,4.

$$f_5(1) = \max\begin{Bmatrix} 4.5 - 1 \\ 5 - 0.5 - 1 \end{Bmatrix} = 3.5, \quad u_5(1) = K \text{或} P.$$

类似地,有

$$f_5(2) = 3, \quad u_5(2) = K \text{或} P;$$

$$f_5(3) = 2.5, \quad u_5(3) = P;$$

$$f_5(4) = 2, \quad u_5(4) = P.$$

当 $k = 4$ 时,公式(4-4)变成

$$f_4(x_4) = \max\begin{cases} r(x_4) - u(x_4) + f_5(x_4 + 1), & u_4 = K; \\ r(0) - u(0) - C(x_4) + f_5(1), & u_4 = P. \end{cases}$$

这时 x_4 的可达状态为1,2,3. 分别代入上式得

$$f_4(1) = 7, \quad u_4(1) = P;$$

$$f_4(2) = 6.5, \quad u_4(2) = P;$$
$$f_4(3) = 6, \quad u_4(3) = P.$$

当 $k = 3$ 时,注意 x_3 的可达状态为 1,2.类似地算出

$$f_3(1) = 10.5, \quad u_3(1) = P;$$
$$f_3(2) = 10, u_3(2) = P.$$

当 $k = 2$ 时,x_2 的可达状态为 1,故

$$f_2(1) = 14, \quad u_2(1) = P.$$

当 $k = 1$ 时,x_1 的可达状态为 0,

$$f_1(0) = 18.5, \quad u_1(0) = K.$$

由此可知,最优更新策略为

$$\{K, P, P, P, K\}$$

或

$$\{K, P, P, P, P\}.$$

最大经济效益为 1.85 万元.

4.4 排序问题

排序问题属组合优化,现已发展成为运筹学的一个独立学科,这里只介绍一些基本问题.

4.4.1 一般提法

这里要介绍的排序问题是指:设有 n 个零件,需要在 m 台设备上加工,如果用 t_{ij} 表示第 i 个零件在第 j 台设备上的加工时间,则称

$$\boldsymbol{T} = (t_{ij})_{n \times m}$$

为加工时间矩阵.在 $\boldsymbol{T}$ 已知的条件下,试问如何安排这 n 个零件在 m 台设备上的加工顺序,才能使得从开始加工第一个零件的时刻起到加工完最后一个零件的时刻止,所用的总时间为最少?

为简便计算,只考虑有 n 个工件需要在 A, B 两台设备上加工,且每个工件都必须经过先 A 后 B 的两道加工工序.已知加工时间矩阵为

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & t_{12} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ \vdots & \vdots \\ a_n & b_n \end{bmatrix}.$$

试问如何安排加工顺序,才能使加工的总时间为最少?

4.4.2 数学模型

将第 k 个工件在两台设备上加工完毕取作第 k 段,$k = 1, 2, \cdots, n$.

用 X 表示等待在 A 上加工的工件集合.对每个 $x \in X$,用 t 表示从在 A 上加工完 x 的时刻算起,直到在 B 上加工完 x 所需的时间.那么 (X, t) 完全可以描述加工

的演变过程，故用(X,t)表示状态变量，并用(X_k,t_k)表示第k段的初始状态．

决策变量u_k取作从第k段的(X_k,t_k)状态出发，在X_k中选取一个工件进行加工．例如取$i\in X_k$，则$u_k=i$．允许决策集合为$D_k=X_k$．

若$u_k=i$，则状态转移方程为

$$X_{k+1}=X_k\setminus\{i\},$$

$$t_{k+1}=\begin{cases}t_k-a_i+b_i, & t_k\geqslant a_i;\\ b_i, & t_k<a_i.\end{cases}$$

用$f_k(X_k,t_k)$表示从状态(X_k,t_k)出发，到将X_k中所有工件加工完毕所用的最少时间；用$f_k(X_k,t_k,i)$表示从状态(X_k,t_k)出发，先在A上加工工件i，然后再对$X_k\setminus\{i\}$中工件采用最优加工顺序所需的时间．由最优化原理，得逆序动态规划方程

$$\begin{cases}f_k(X_k,t_k)=\min\limits_{i\in X_k}\begin{cases}a_i+f_{k+1}(X_k\setminus\{i\},t_k-a_i+b_i), & t_k\geqslant a_i;\\ a_i+f_{k+1}(X_k\setminus\{i\},b_i), & t_k<a_i,\end{cases}\\ f_{n+1}(X_{n+1},t_{n+1})=0,k=n,\cdots,2,1.\end{cases}$$

4.4.3 最优排序程序

定理2 使

$$f_k(X_k,t_k,i,j)\leqslant f_k(X_k,t_k,j,i)$$

的充要条件是

$$\min\{a_i,b_j\}\leqslant\min\{a_j,b_i\}. \tag{4-5}$$

由定理2知，在第k段，将工件i排在工件j之前加工的充要条件是(4-5)式成立．

根据定理2，可以得下述最优排序程序：

1° 令

$$c=\min\{a_1,a_2,\cdots,a_n;b_1,b_2,\cdots,b_n\}.$$

2° 若$c=a_i$，则把工件i排在第一位，并从工件集合中去掉它．

3° 若$c=b_i$，则把工件i排在最后一位，并从工件集合中去掉它．

4° 对剩下的工件集合重复上述步骤，直到工件集合为空集为止．

例4 已知5个工件，它们在2台设备上的加工时间矩阵为

$$T=\begin{bmatrix}3 & 2 & 5 & 6 & 4\\ 7 & 9 & 8 & 3 & 6\end{bmatrix}^{\mathrm{T}},$$

试求最优排序．

解 按最优排序程序，易知最优排序为

$$2\to1\to5\to3\to4.$$

5 随机动态规划简介

确定型动态规划是指多阶段决策问题的过程是确定的，即选定策略$\{u_1,u_2,$

$\cdots,u_n\}$ 后,过程的状态序列 $\{x_1,x_2,\cdots,x_{n+1}\}$ 也就随之确定.

随机型动态规划是指多阶段决策问题的过程不是确定的,而是随机的,即给定策略后,过程的状态序列 $\{x_1,x_2,\cdots,x_n,x_{n+1}\}$ 是随机的,也即 $\{x_1,x_2,\cdots,x_n,x_{n+1}\}$ 形成了一个具有某种概率结构的离散参数的马尔可夫过程.

对于随机型多阶段决策问题,从数学上可以证明,最优化原理仍然成立.

5.1　随机道路问题

5.1.1　简单随机道路问题的基本解法

图5-1是一个简单的网络,其中线段(称为弧)上的数字是通过该弧的费用.在每个结点上都有两个决策可选,对角向上(记作 U)或对角向下(记作 D).试问沿哪条道路由 A 行进到直线 B,才能使总费用最小?

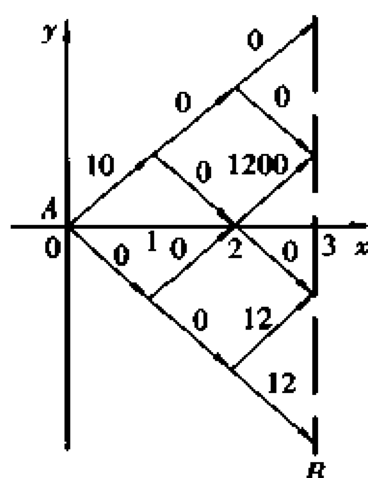


图5-1

如果是确定型的,就是第1章1.1节中的最短路问题.**随机道路问题**是指,当一个旅行者在某个结点被通知向上走时,他只以3/4的概率记住 U,而以1/4的概率忘掉 U 向下走;反之,如果在这个结点被通知向下走时,他同样以3/4的概率记住 D,而以1/4的概率忘掉 D 向上走.旅行者在每个结点都遵守这个规律,而不管在前面的结点是否接到这个通知.这样,只能求出他经过各种可能道路的概率.随机道路问题的求解就是要求这样的路线:其相应的费用的期望值最小.

由于对随机道路来说,用决策序列控制过程和用策略控制过程所得到的解是不同的,因此有两种形式的解.称由决策序列确定的解为**开环控制**,而由策略确定的解为**反馈控制**.

1. 用穷举法求最优开环控制

图5-1给出的道路问题,在确定型中,每个决策序列由3个决策构成,总共有8个决策序列,它们是

		上			上
	上			上	
		下			下
上			下		
		上			上
	下			下	
		下,			下.

在随机问题中,若决策序列是 $D—U—D$,则以27/64的概率(即三个决策 D,U,D 都记住)产生一条道路:

下 — 上 — 下,费用为0;

以9/64的概率(即记住2个决策,忘掉1个决策)产生三条道路:

下一上一上， 费用为 1 200，
下一下一下， 费用为 12，
上一上一下， 费用为 10；

以 3/64 的概率(即记住 1 个决策，忘掉 2 个决策)产生三条道路：

下一下一上， 费用为 12，
上一上一上， 费用为 10，
上一下一下， 费用为 10；

以 1/64 的概率(即 3 个决策都忘掉)产生一条道路：

上一下一上， 费用为 1210.

将以上 8 个费用分别乘以它们相应的概率，然后再相加，便得决策序列 $D—U—D$ 的期望费用

$$
\begin{aligned}
E_{DUD} &= 27/64 \times 0 + 9/64 \times (1200 + 12 + 10) + \\
&\quad 3/64 \times (12 + 10 + 10) + 1/64 \times 1210 \\
&= 192\frac{1}{4}.
\end{aligned}
$$

这样的决策序列共有 8 个，按同样的计算方法，分别计算出它们的期望值，列入表 5-1 中．其中最小者为 $120\frac{3}{4}$，即决策序列 $U—U—D$ 具有最小期望费用，而在确定型中，决策序列 $D—U—D$ 费用最小(为 0).

表 5-1

决策序列	期望费用
$D—U—D$	$192\frac{1}{4}$
$D—U—U$	$567\frac{1}{4}$
$D—D—U$	$346\frac{3}{4}$
$D—D—D$	$121\frac{3}{4}$
$U—D—U$	$572\frac{1}{4}$
$U—D—D$	$197\frac{1}{4}$
$U—U—D$	$120\frac{3}{4}$
$U—U—U$	$345\frac{29}{32}$

2. 用动态规划方法求最优反馈控制

用(x,y)(图 5-1 中的结点)表示状态变量．决策变量取 U 或取 D. 如果取 U，则以 3/4 的概率转到状态$(x+1,y+1)$，这时设费用为 $a_u(x,y)$；或以 1/4 的概率转到状态$(x+1,y-1)$，这时设费用为 $a_d(x,y)$. 如果取 D，则以 3/4 的概率转到状态$(x+1,y-1)$，这时费用为 $a_d(x,y)$；或以 1/4 的概率转到状态$(x+1,y+1)$，这时费用为 $a_u(x,y)$.

定义 1 最优期望函数 $S(x,y)$ 为从状态(x,y)出发，采用最优反馈控制策略时，其余过程的期望费用．

于是，由最优化原理的随机形式，便得逆序随机动态规划模型，即递推关系为

$$
S(x,y) = \min\begin{cases} \frac{3}{4}[a_u(x,y) + S(x+1,y+1)] + \\ \frac{1}{4}[a_d(x,y) + S(x+1,y-1)], \quad \text{取 } U; \\ \frac{1}{4}[a_u(x,y) + S(x+1,y+1)] + \\ \frac{3}{4}[a_d(x,y) + S(x+1,y-1)], \quad \text{取 } D, \end{cases} \tag{5-1}
$$

边界条件为

$$
S(3,3) = S(3,1) = S(3,-1) = S(3,-3) = 0. \tag{5-2}
$$

例1　利用递推公式(5-1)及(5-2)求解图5-1中的随机道路问题.

解　当 $x=2,y=2,0,-2$ 时,

$$S(2,2)=\min\left\{\frac{3}{4}(0+0)+\frac{1}{4}(0+0),\frac{1}{4}(0+0)+\frac{3}{4}(0+0)\right\}=0,$$

$$p(2,2)=U\text{或}D.$$

类似可求出

$$S(2,0)=300,\quad p(2,0)=D.$$

$$S(2,-2)=12,\quad p(2,-2)=U\text{或}D.$$

当 $x=1,y=1,-1$ 时,

$$S(1,1)=75,\quad p(1,1)=U.$$

$$S(1,-1)=84,\quad p(1,-1)=D.$$

当 $x=0,y=0$ 时,

$$S(0,0)=84\frac{1}{4},\quad p(0,0)=D.$$

即最优期望费用为 $84\frac{1}{4}$.再由反向追踪可知最优反馈策略为

$$\{D,D,U\}\quad\text{或}\quad\{D,D,D\},$$

可见期望费用比用决策序列得到的期望费用小.这是因为最优反馈控制利用了在每一状态的所有信息,因此通常要比最优开环控制产生的期望费用小.

5.1.2　随机停止时间问题

在很多问题中,不仅决策转移是随机的,而且过程的阶段也是随机的.例如生命的阶段,一场排球赛的阶段(局)等都是随机的.这样的问题称为**随机停止时间问题**.

如图5-2所示,假设这个随机道路问题的阶段不定,即当旅行者从 A 点出发后,并不知道是到达直线 C 还是到达直线 D 才停止前进.已知到达 C 停止时的概率是 p_C,到达 D 停止时的概率是 $p_D=1-p_C$.当旅行者到达直线 B,而且尚未作出决策前,并不知道过程到哪一条直线结束.

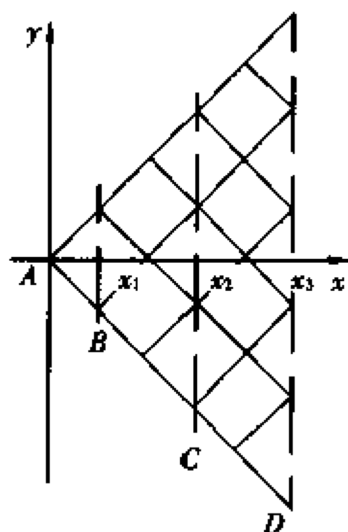


图5-2

假设已处在过程的某一阶段.若阶段数 x 满足 $x_1\leqslant x\leqslant x_2$,则在 x_1 阶段已经知道被通知在什么地方停止;若 $x<x_1$,当然还不知道这个信息;若 $x>x_2$,则 x_3 就是停止线.因此,在不同阶段,其最优期望值函数是不同的.规定如下:

(1) 当 $x\leqslant x_1-1$ 时,用 $F(x,y)$ 表示从状态 (x,y) 出发,其余过程的最小期望费用;

(2) 当 $x_1\leqslant x\leqslant x_2-1$ 时,用 $G(x,y,z)$ 表示从状态 (x,y) 出发,且在 z 处($z=x_2$ 或 $z=x_3$)停止时,其余过程的最小费用;

(3) 当 $x \geq x_2$ 时,用 $H(x,y)$ 表示从状态(x,y)出发,且在 $x = x_3$ 停止时,其余过程的最小费用．

实际上,这里只有第1种情况是期望费用,因为一旦第2或第3两个最优函数的自变量给定后,过程的阶段就知道了．由此,便得到如下的递推关系,也即这种问题的逆序动态规划模型:

1° 当 $x \leq x_1 - 1$ 时,

$$F(x,y) = \min\begin{Bmatrix} a_u(x,y) + F(x+1,y+1), \\ a_d(x,y) + F(x+1,y-1) \end{Bmatrix},$$

边界条件为

$$F(x_1-1,y) = \min\begin{Bmatrix} a_u(x-1,y) + p_C G(x_1,y+1,x_2) + p_D G(x_1,y+1,x_3), \\ a_d(x-1,y) + p_C G(x_1,y-1,x_2) + p_D G(x_1,y-1,x_3) \end{Bmatrix}.$$

2° 当 $x_1 \leq x \leq x_2 - 1$ 时,

$$G(x,y,z) = \min\begin{Bmatrix} a_u(x,y) + G(x+1,y+1,z), \\ a_d(x,y) + G(x+1,y-1,z) \end{Bmatrix},$$

边界条件为:对所有的 y,有

$$G(x_2,y,x_2) = 0, \quad G(x_2,y,x_3) = H(x_2,y).$$

3° 当 $x \geq x_2$ 时,

$$H(x,y) = \min\begin{Bmatrix} a_u(x,y) + H(x+1,y+1), \\ a_d(x,y) + H(x+1,y-1) \end{Bmatrix},$$

边界条件为:对所有的 y,有

$$H(x_3,y) = 0.$$

计算次序是从 x_3 逆序进行:先计算 H,再计算 G(对两种停止情况的每一种,它都给出在每个结点的不同费用和可能的不同决策),最后计算出 F,即为所求．

最后必须指出,有些实际问题,在做出决策和完成这个决策之间要经过一段时间．例如,在库存问题中,订货和到货就须经过一段时间．这种问题称为**时间延迟问题**．这里不再讨论．

5.2　随机设备更新问题

5.2.1　什么是随机设备更新问题

随机设备更新问题与第4章4.3节中的确定型设备更新问题有很多区别,这里只考虑两点主要区别:一是设备使用费用与设备役令的关系不是确定的,而是一个随机变量;二是设备在任何1年的年底都可能遭到灾难性破坏而必须更新,不像在确定型问题中只考虑常规老化而更新．

在确定型问题中,是以最大效益为目标建立的更新模型．现在将换一个角度,以最小费用为目标建立更新模型．为此,先规定几个记号:

用 $u(i,j)$ 表示年初役令为 i 的一台设备在1年中的纯操作费用是 $j \in \{0,1,$

$\cdots,J\}$ 的概率.

用 $C(i)$ 表示役令为 i 的一台设备的更新费用函数(与确定型问题中 $C(i)$ 一样).

用 $v(i)$ 表示役令刚变为 i 的一台设备处于损坏状态的折旧费.

用 $q(i)$ 表示年初役令为 i 的正在工作的一台设备,在年末时处于损坏状态的概率.

$S(i)$ 为 $n+1$ 年年初役令为 i 的一台设备处于工作状态的利用值.

$T(i)$ 为 $n+1$ 年年初役令为 i 的一台设备处于损坏状态的利用值.

p 表示购买一台新设备的价格.

m 表示过程开始时设备的役令.

5.2.2 动态规划模型

按设备使用的年限(包括使用旧设备和更新后的新设备)把问题分成 n 个阶段.

状态变量 x_k 取作从过程开始时起,设备的役令 i,即 $x_k = i$.

决策变量 u_k 取作继续使用旧设备(记作 K),或者使用更新后的新设备(记作 P). 于是,状态转移方程为

$$x_{k+1} = \begin{cases} 1, & u_k = P; \\ i+1, & u_k = K. \end{cases}$$

阶段指标函数为

$$d(x_k, u_k) = \begin{cases} C(i) + \sum_{j=0}^{J} ju(0,j), & u_k = P; \\ \sum_{j=0}^{J} ju(i,j), & u_k = K. \end{cases}$$

用 $f_k(i)$ 表示从第 k 年年初开始,使用一台役令 $x_k = i$ 的设备,到第 n 年年末时过程的最小期望费用. 由最优化原理,得随机设备更新问题的逆序动态规划模型

$$f_k(i) = \min\begin{cases} C(i) + \sum_{j=0}^{J} ju(0,j) + q(0)[p - v(1) + \\ f_{k+1}(0)] + (1-q(0))f_{k+1}(1), u_k = P, i = 1,2,\cdots,k-1; \\ \sum_{j=0}^{J} ju(i,j) + q(i)[p - v(i+1) + \\ f_{k+1}(0)] + (1-q(i))f_{k+1}(i+1), u_k = K, i = m+k-1. \end{cases}$$

其中 $k = n-1,\cdots,2,1$.

$$f_k(0) = \sum_{j=0}^{J} ju(0,j) + q(0)[p - v(1) + f_{k+1}(0)] + (1-q(0))f_{k+1}(1).$$

边界条件为

$$f_n(i)=\min\begin{cases}C(i)+\sum_{j=0}^{J}ju(0,j)-q(0)T(1)-(1-q(0))S(1), & u_k=P;\\ \sum_{j=0}^{J}ju(i,j)-q(i)T(i+1)-(1-q(i))S(i+1), & u_k=K.\end{cases}$$

5.3 库存问题

5.3.1 动态库存问题及其基本术语

假设某企业在连续 n 个阶段里专门供应某种产品，产品的初始库存为 a 件，每一阶段初，企业采购若干产品，同时供应社会若干产品，到阶段末规定库存为 b 件，试问各阶段应采购多少件产品，才能使总费用最小?这类问题称为**动态库存问题**．

下面介绍一些有关的常用的基本术语．

阶段：按时间划分，可以是天、月或年，一般各阶段是等长的．

计划水平：一个问题中的阶段数．

库存水平：库存的实有数．

采购：分购买(决定采购立即得到产品)，订购(采购后，要经过 $\lambda>0$ 个阶段才能得到产品)和随机订购(交货期为具有一定概率分布的随机变量)三种情形．其中 λ 称为交货延期．

社会需求：分确定型需求(每一阶段的需求数量是确定的)和随机型需求(在 i 阶段需求 d 件产品的概率为 $p(d)$，且各阶段的需求是彼此独立的随机变量)．

供求关系：设 x_i 和 y_i 分别是 i 阶段订货前和订货后(包括已订货但还未到货)的库存量，z_i 是 i 阶段的订货件数，交货延期为 λ，则 $y_i=x_i+z_i$．再设 w_i 是 $i-\lambda$ 阶段订货 $z_{i-\lambda}$，在 i 阶段已交货，但未产生需求前的库存量(包括已订货而未交货者)，即 w_i 是 i 阶段实际可达的库存量．当 $\lambda=0$ 时，$w_i=y_i$；当 $\lambda>0$ 时，$w_{i+\lambda}=y_i-(d_i+d_{i+1}+\cdots+d_{i+\lambda-1})$，其中 d_i 是 i 阶段的实际产品需求．

为确定 x_{i+1}，应考虑 $d_i>w_i$ 的可能．这时若欠缺的 d_i-w_i 件产品延期供应，则称为允许缺货；若欠缺的 d_i-w_i 件产品放弃供应，则称失销．在失销阶段，若 $\lambda=0$，则 $x_{i+1}=\max\{y_i-d_i,0\}$；若 $\lambda>0$，则情况较复杂，这里不再说明．

订货费用：用 $C(z)$ 表示 i 阶段订购 z 件产品的订货费．它可在 i 阶段支付，也可在 $i+\lambda$ 阶段支付．

存储缺货费用：当 $w_i>d_i$ 时，净存产品 w_i-d_i 件，设存储费用为 $h(w_i-d_i)$，一般包括仓库租用费、保险费、税金、维修、老化及资金积压的利息等．当 $w_i<d_i$ 时，则产生两项费用，即缺货费 $\pi_i(d_i-w_i)$ 和(空)存储费 $h(0)$．用 $L_i^0(w_i,d_i)$ 表示存储缺货费用，则

$$L_i^0(w_i,d_i)=\begin{cases}h_i(w_i-d_i), & w_i>d_i;\\ \pi_i(d_i-w_i)+h_i(0), & w_i<d_i.\end{cases}\tag{5-3}$$

对随机情形，如设 i 阶段需求为 d 的概率为 $p_i(d)$，则期望存储缺货费为

$$L_i^0(w_i) = \sum_{d=0}^{\infty} L_i^0(w_i, d) p_i(d). \tag{5-4}$$

终结费用:在 n 阶段末,若库存恰有 b 件产品,则正好满足要求.若库存为 $u \neq b$ 时,有两种可能:如 $u > b$,则应卖掉 $v = u - b$ 件产品,得退货费 $-S(v)$;如 $u < b$,且允许缺货,则应再进货 $v = u - b$ 件产品,发生费用 $b(v)$.将这两项费用统称为终结费用,记作 $t(v)$,即

$$t(v) = \begin{cases} -S(v), & v \geqslant 0; \\ b(-v), & v < 0. \end{cases} \tag{5-5}$$

折合因子:假设每一阶段资金的利率为 r,即若 i 阶段初资金为 e,则在 $i+k$ 阶段初资金为 $e(1+r)^k$,或者说折合到 $i+k$ 阶段资金为 $e(1+r)^k$.当 $r>0, k<0$ 时,$(1+r)^k < 1$.如记 $1+r=\beta, \alpha = 1/\beta$,则 i 阶段初的资金 e 折合到1阶段初为 $e\alpha^{i-1}$,折合到 n 阶段末为 $e\beta^{n-i+1}$.这里 α, β 称为折合因子.当 $\alpha = \beta = 1$ 时,资金与时间无关.

概率卷积:设 $D_i = d$ 的概率为 $p_i(d)(i = 1,2,\cdots,n)$,D_i 为彼此独立的随机变量,则

$$D_i + D_{i+1} + \cdots + D_j = d$$
$$(i = 1,2,\cdots,n; j = i, i+1,\cdots,n)$$

的概率叫做 $p_i(d), p_{i+1}(d), \cdots, p_j(d)$ 的卷积,记作 $p_{i,j}(d)(d = 0,1,\cdots)$,并可用下面公式计算:

$$\begin{cases} p_{i,i}(d) \equiv p_i(d), \\ p_{i,k}(d) = \sum_{l=0}^{d} p_{i,k-1}(l) p_k(d-l), k = i+1,\cdots,j. \end{cases} \tag{5-6}$$

5.3.2 不延期交货($\lambda = 0$)的 n 阶段库存问题

1. 确定型

静态模型　假设 $\alpha = 1$,且每个阶段 i 的需求 d_i 已知,初始和终结库存均为0,不允许缺货,试问每阶段订货量各多少,才能使总费用最小?

设 i 阶段的订货量为 z_i,i 阶段末的库存水平为 v_i.由于不许缺货,故

$$v_i = \sum_{j=1}^{i} z_j - \sum_{j=1}^{i} d_j \geqslant 0, \quad i = 1,2,\cdots,n.$$

令 $z = (z_1, z_2, \cdots, z_n)$,则可行订货策略集合为

$$S = \{z \mid z_i \geqslant 0, v_i \geqslant 0, v_n = 0, i = 1,2,\cdots,n\}.$$

再用 $C_i(z_i)$ 表示 i 阶段订货 z_i 件的费用,$h_i(v_i)$ 表示 i 阶段末库存为 v_i 时的存储费用.由于不许缺货($v_i \geqslant 0$),故终结费用为

$$t(v_n) = -S(v_n).$$

于是,问题变成

$$\min_{z \in S} J(z) = \sum_{i=1}^{n} C_i(z_i) + \sum_{i=1}^{n} h_i\left(\sum_{j=1}^{i}(z_j - d_j)\right).$$

这就是不延期交货、确定型 n 阶段库存问题的静态模型 .

动态模型　假设将从 1 到 k 阶段的库存水平 v 取作状态变量，每个阶段的订货件数 z 取作决策变量 . 由此可知，决策集合、状态转移方程、阶段指标依次为

$$D_k = \{z \mid 0 \leqslant z \leqslant v + d_k\},$$

$$\tilde{v} = v - z + d_k,$$
$$d(v,z) = C_k(z) + h_k(v).$$

用 $f_k(v)$ 表示在 k 阶段末必须有 v 件库存时，从 1 到 k 阶段的最小费用，根据最优化原理，得顺序递推公式

$$\begin{cases} f_k(v) = \min\limits_{z \in D_k}\{C_k(z) + h_k(v) + f_{k-1}(v - z - d_k)\}, \\ f_1(v) = C_1(v + d_1) + h_1(v), k = 2,3,\cdots,n. \end{cases} \tag{5-7}$$

对每一个 k，应在$[0, \sum\limits_{j=k+1}^{n} d_j]$ 上的每个正整数 v 处计算$f_i(v)$，最后 $f_n(0)$ 即为所求 .

2. 随机型

静态模型　假设 $\alpha \neq 1$，第 k 阶段开始时库存为 x，第 k 阶段需求为 d 的概率为 $p_k(d)$，在允许缺货的情况下，求最小期望费用 .

设第 k 阶段订货 $y - x$ 件，由于 $\lambda = 0$ 没有延期交货，故有 y 件产品可供需求，即实际库存量为 $w = y$. 于是，在第 k 阶段的订货费为 $C_k(y - x)$，第 k 阶段有 w 件产品，需求为 d 的存储短缺费，期望存储短缺费及终结费用依次由(5-3) 式、(5-4) 式及(5-5) 式确定 . 由此可知，所论问题变成

$$\min_{y \geqslant x}\left[\sum_{k=1}^{n} C_k(y - x) + \sum_{k=1}^{n} L_k^0(y)\right],$$

其中 $L_k^0(y)$ 由(5-4) 及(5-3) 式确定 . 这就是不延期交货，可以缺货，随机型 n 阶段库存问题的静态模型 .

动态模型　用 $f_k(x)$ 表示第 k 阶段开始时库存为 x，且采用最优订货策略，从第 k 到第 n 阶段的期望总折合费用，则易知此问题的逆序动态规划模型为

$$\begin{cases} f_k(x) = \min\limits_{y \geqslant x}\{C_k(y - x) + L_k^0(y) + \alpha \sum\limits_{d=0}^{\infty} f_{k+1}(y - d) p_{k+1}(d)\}, \\ f_{n+1}(x) = t(x), k = n,\cdots,2,1. \end{cases} \tag{5-8}$$

实际计算时，必须对每个可能的 x 计算$f_k(x)$，若 $y_k(x) - k$ 使(5-8) 式成立，则 $y_k(x) = x_k$(某个 x)，就是最优订货量 .

例 2　试计算下面问题的最优期望费用 . 假设 $n = 3, x_1 = 0, \alpha = 1$，且

$$p(0) = 1/4, p(1) = 1/2, p(2) = 1/4;$$
$$C(0) = 0, C(1) = 3, C(2) = 5,$$
$$C(3) = 6, z \geqslant 4 \text{时}, C(z) = \infty;$$

又 $v \geqslant 0$ 时，

$$h(v) = v, \quad \pi(v) = 5v, \quad S(v) = 2v, \quad b(v) = 4v.$$

解　不难知道，此问题的最大可能库存为 9，而最大可能允许缺货为 6，故 $x =$

$0,1,\cdots,9$ 及 $x=-1,-2,\cdots,-6$. 由(5-5)式知

$$t(v)=\begin{cases}-2v, & v\geqslant 0;\\ -4v, & v<0.\end{cases}$$

再由(5-8)式的边界条件知:当 $x=0,1,\cdots,9$ 时, $f_4(x)=-2x$;当 $x=-1,-2,\cdots,-6$ 时, $f_4(x)=-4x$. 由此,按递推公式(5-8)便可计算 $f_3(x)$. 但须注意,当库存量 x 超过其余过程的最大需求时决不会订货,因此只须对 $x=6,\cdots,1,0,-1,\cdots,-4$ 计算 $f_3(x)$, 于是

$$f_3(6)=\min_{6\leqslant y\leqslant 9}\left\{C(y-6)+L^0(y)+\sum_{d=0}^{2}f_4(y-d)p(d)\right\}.$$

由于 $y=6,7,8,9$ 时, $C(0)=0, C(1)=3, C(2)=5, C(3)=6$;又

$$L^0(y,d)=\begin{cases}y-d, & y\geqslant d;\\ 5(y-d), & y<d,\end{cases}$$

而 $y\geqslant 6, d=0,1,2$, 故 $L^0(y,d)=y-d$. 所以

$$L^0(y)=\sum_{d=0}^{2}(y-d)p(d).$$

当 $y=6,7,8,9$ 时,

$$L^0(6)=5,\quad L^0(7)=6,\quad L^0(8)=7,\quad L^0(9)=8.$$

最后再注意 $y=6,7,8,9$ 时,

$$\sum_{d=0}^{2}f_4(y-d)p(d)=2(1-y),$$

即依次为 $-10,-12,-14,-16$, 于是

$$f_3(x)=\min\{0+5-10, 3+6-12, 5+7-14, 6+8-16\}=-5,$$

$$y_3(6)=6.$$

类似可得

$$\begin{aligned}&f_3(5)=-4, y_3(5)=5; f_3(4)=-3, y_3(4)=4;\\ &f_3(3)=-2, y_3(3)=3; f_3(2)=-1, y_3(2)=2;\\ &f_3(1)=2, y_3(1)=1\text{ 或 }2; f_3(0)=4, y_3(0)=2\text{ 或 }3;\\ &f_3(-1)=5, y_3(-1)=2; f_3(-2)=8, y_3(-2)=1;\\ &f_3(-3)=15, y_3(-3)=0; f_3(-4)=24, y_3(-4)=-1.\end{aligned}$$

按同样方法计算 $f_2(x)$, 这时 $x=3,2,1,0,-1,-2$, 于是得

$$\begin{aligned}f_2(3)&=\min_{3\leqslant y\leqslant 6}\left\{C(y-3)+L^0(y)+\sum_{d=0}^{2}f_3(y-d)p(d)\right\}\\ &=\min\left\{0+2-\frac{1}{2}, 3+3-2, 5+4-3, 5+6-4\right\}=1/2,\end{aligned}$$

$$y_2(3)=3;$$

$$f_2(2)=2\frac{2}{4}, y_2(2)=2; f_2(1)=5\frac{1}{4}, y_2(1)=1;$$

$$f_2(0)=7\frac{1}{2}, y_2(0)=3; f_2(-1)=8\frac{3}{4}, y_2(-1)=2;$$

$$f_2(-2)=11\frac{1}{4},\quad y_2(-2)=1.$$

最后,得

$$\begin{aligned}f_1(0)&=\min_{0\leqslant y\leqslant 3}\{C(y)+L^0(y)+\sum_{d=0}^{2}f_2(y-d)p(d)\}\\&=\min\{0+5+9\frac{1}{16},3+1\frac{1}{2}+7\frac{1}{4},5+1+5\frac{3}{16},6+2+3\frac{1}{16}\}\\&=11\frac{1}{16},\end{aligned}$$

$$y_1(0)=3.$$

由上述计算可知,1 阶段开始时购买 3 件产品是最优的,这时最小期望费用为 $11\frac{1}{16}$.

5.3.3 延期交货($\lambda>0$)的 n 阶段库存问题

1. 确定型

只考虑动态模型的情况．在允许缺货时,本阶段的问题是5.3.2小节中确定型动态模型(5-7)式的推广．在 $\lambda=0$ 的模型中每阶段购买产品后,必须保证充分供应需求,而在 $\lambda>0$ 的模型中可以发生缺货,且不足部分将在问题结束之前供应完毕.因此,求解 $\lambda>0$ 的模型只须对 $\lambda=0$ 的模型(5-7)式作两点修改．

第一,因为在每个阶段不需要保证供应,所以订购量可以从0到 σ(σ 为订货上限),在第 i 阶段末,库存水平可以是负整数,它介于 $-\sum_{j=1}^{i}d_j$ 与 $i\sigma-\sum_{j=1}^{d}d_j$ 之间．

第二,把存储费用 $h_i(v_i)$ 改为存储缺货费用 $L_i^0(w_i,d_i)$.

2. 随机型

首先注意,当 $\lambda=1$ 时,用 $\lambda=0$ 时的公式很容易解决,因此只讨论 $\lambda\geqslant 2$ 的情形.

这时在 i 阶段初订购 z 件产品且将在 $i+\lambda$ 阶段初到货,并假设订货费在实际交货时支付,而终结费用用 $f_{n+1}(x)=t(x)$ 或 $-S(x)(x\geqslant 0)$ 给出,这要看是允许缺货还是失销而定．当然,在 $n-\lambda+1,n-\lambda+2,\cdots,n-\lambda$ 阶段绝不会订货,因为交货时过程已结束．下面给出一种逆序动态规划解法．

使用5.3.1中的符号,分别用 x_i 和 y_i 表示第 i 阶段订货前后(包括已订货还未交货)的库存量,z_i 为第 i 阶段的订购件数,$w_{i+\lambda}$ 为 $i+\lambda$ 阶段的实际库存量(不包括已定货而未交货),d_i 为第 i 阶段的社会需求,则有

$$y_i=x_i+z_i,$$

$$w_{i+\lambda}=y_i-\sum_{j=i}^{i+\lambda-1}d_j.$$

为简单计,以下略去 x_i,y_i,z_i,d_i 等的下标 i. 注意 $L_i^0(y)$ 表示 $y_i=y$ 时第 $i+\lambda$ 阶段的期望库存费用,故易知

$$L_i^0(y) = \sum_{d=0}^{\infty}\left[\sum_{d'=0}^{\infty}L_{i+\lambda}^0(y-d,d')p_{i+\lambda}(d')\right]p_{i,i+\lambda-1}(d)$$
$$= \sum_{d=0}^{\infty}L_{i+\lambda}^0(y-d)p_{i,i+\lambda-1}(d),$$

其中 $L_{i+\lambda}^0(y-d,d')$，$L_{i+\lambda}^0(y-d)$，$p_{i,i+\lambda-1}(d)$ 分别由(5-3)、(5-4)、(5-6)式确定.

取 x 为状态变量，y 为决策变量，则状态转移方程、决策集合及阶段指标函数分别为

$$\tilde{x} = y - d,\quad y \geqslant x,\quad C_i(y-x) + L_i^0(y).$$

由于第 i 阶段订购 z 件产品的费用 $C(z)$ 要在 $i+\lambda$ 阶段支付，而不影响从 i 到 $i+\lambda-1$ 阶段的支付费用，从而不影响从 i 到 $i+\lambda-1$ 阶段的期望存储费用，因此在最优函数定义中可以不考虑 i 到 $i+\lambda-1$ 阶段的费用．于是，可用 $f_i(x)$ 表示从状态 x 出发，采用最优策略，从第 $i+\lambda$ 阶段到第 n 阶段的最小期望费用对 $i+\lambda$ 阶段的总折合值，则由最优化原理，便得此问题的逆序动态规划模型为

$$\begin{cases} f_i(x) = \min\limits_{y\geqslant x}\{C_i(y-x) + L_i^0(y) + \alpha\sum\limits_{d=0}^{\infty}f_{i+1}(y-d)p_i(d)\}, \\ f_{n-\lambda+1}(x) = \sum\limits_{d=0}^{\infty}t(x-d)p_{n-\lambda+1,n}(d), i = n-\lambda,\cdots,2,1. \end{cases} \tag{5-9}$$

利用递推公式(5-9)逐次求解，最后得到 $f_1(x)$，它是从 $\lambda+1$ 阶段到 n 阶段的最小期望费用对 $\lambda+1$ 阶段的总折合值，因此它自然不包括从 1 阶段到 λ 阶段的期望存储缺货费用．

由于 1 阶段到 λ 阶段的期望存储缺货费用对 1 阶段的折合值为

$$L_1^0(x_1) + \sum_{i=2}^{\lambda}\left[\alpha^{i-1}\sum_{d=0}^{\infty}L_i^0(x_1-d)p_{1,i-1}(d)\right],$$

所以从状态 x_1 出发，采用最优策略，从第 1 阶段到第 n 阶段的最小期望总折合费用（对 1 阶段的折合）为

$$C = L_1^0(x_1) + \sum_{i=2}^{\lambda}\left[\alpha^{i-1}\sum_{d=0}^{\infty}L_i^0(x_1-d)p_{1,i-1}(d)\right] + \alpha^{\lambda}f_1(x_1). \tag{5-10}$$

例 3 试求下述库存问题的最优订货策略及最小期望费用．

$$n = 4, \lambda = 2, \alpha = 1, x_1 = 0;$$
$$p(0) = 3/4, p(1) = 1/4;$$
$$C(0) = 0, C(1) = 3, \text{当 } z \geqslant 2 \text{ 时}, C(z) = \infty;$$
$$\text{当 } v \geqslant 0 \text{ 时}, h(v) = v, \pi(v) = 3v, S(v) = 2v, b(v) = 4v.$$

解 首先注意，这个问题的最大库存量为 2，最大缺货量为 -2.

由(5-3)式及(5-4)式，有

$$L^0(w,d) = \begin{cases} w-d, & w \geqslant d; \\ 3(d-w), & w < d. \end{cases}$$

$$L^0(2) = \sum_{d=0}^{\infty}L^0(2,d)p(d) = L^0(2,0)p(0) + L^0(2,1)p(1)$$

$$= 2 \times \frac{3}{4} + 1 \times \frac{1}{4} = \frac{7}{4}.$$

同理可得

$$L^0(1) = \frac{3}{4}, L^0(0) = \frac{3}{4}, L^0(-1) = \frac{15}{4}, L^0(-2) = \frac{27}{4}, L^0(-3) = \frac{39}{4}.$$

再由公式(5-6)得

$$p_{i,i+1}(0) = p_{i,i}(0)p_{i+1}(0) = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16},$$

$$p_{i,i+1}(1) = p_{i,i}(0)p_{i+1}(1) + p_{i,i}(1)p_{i+1}(0)$$

$$= \frac{3}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{3}{4} = \frac{6}{16},$$

$$p_{i,i+1}(2) = p_{i,i}(0)p_{i+1}(2) + p_{i,i}(1)p_{i+1}(1) + p_{i,i}(2)p_{i+1}(0)$$

$$= \frac{3}{4} \times 0 + \frac{1}{4} \times \frac{1}{4} + 0 \times \frac{3}{4} = \frac{1}{16}.$$

于是,由 $L_i(y) = \sum_{d=0}^{\infty} L_{i+\lambda}^0(y-d)p_{i,i+\lambda-1}(d)$ 及上述结果得

$$L_2(2) = L^0(2)p_{2,3}(0) + L^0(1)p_{2,3}(1) + L^0(0)p_{2,3}(2) = \frac{21}{16}.$$

类似得

$$L_2(1) = \frac{15}{16}, L_2(0) = \frac{9}{4}, L_2(-1) = \frac{21}{4}, L_2(-2) = \frac{33}{4}.$$

最后,由公式(5-5)有

$$t(3) = -6, t(2) = -4, t(1) = -2, t(0) = 0,$$

$$t(-1) = 4, t(-2) = 8, t(-3) = 12, t(-4) = 16.$$

现在,利用递推公式(5-9),先计算边界条件

$$f_3(2) = t(2)p_{3,4}(0) + t(1)p_{3,4}(1) + t(0)p_{3,4}(2) = -3,$$

$$f_3(1) = -\frac{7}{8}, f_3(0) = 2, f_3(-1) = 6, f_3(-2) = 10.$$

再计算 $f_2(x)$. 因为 $x_1 = 0$, 且仅在 1、2 阶段订货,每次订货量为 0 或 1,故 $-1 \leqslant x \leqslant 1$. 又因为这时 $y = 0,1,2$, 即 $x = 1$ 时, $y = 1$ 或 2; $x = 0$ 时, $y = 0$ 或 1; $x = -1$ 时, $y = -1$ 或 0, 所以

$$f_2(1) = \min_{1 \leqslant y \leqslant 2} \{C_2(y-1) + L_2(y) + \sum_{d=0}^{1} f_3(y-d)p_2(d)\} = \frac{25}{32}, y_2(1) = 1;$$

$$f_2(0) = \min\{\frac{21}{4}, \frac{121}{32}\} = \frac{21}{4}, \quad y_2(0) = 1;$$

$$f_2(-1) = \min\{\frac{49}{4}, \frac{33}{4}\} = \frac{33}{4}, \quad y_2(-1) = 0.$$

再计算 $f_1(x)$. 因这时 $x_1 = 0$, 故

$$f_1(0) = \min_{0 \leqslant y \leqslant 1} \{C_1(y-0) + L_1(y) + \sum_{d=0}^{1} f_2(y-d)p_1(d)\}$$

$$= 5\frac{15}{32}, y_1(0) = 1.$$

最后，由(5-10)式得从1到4阶段的最小期望存储缺货费为

$$L_1^0(0) + \left[\frac{3}{4}L_2^0(2) + \frac{1}{4}L_2^0(-1)\right] + f_1(0)$$

$$= 2\frac{8}{32} + 5\frac{15}{32} = 7\frac{23}{32}.$$

最优订货策略由 $y_1(0) = 1$ 决定，即 $y_1(0) - x_1 = 1$，亦即第1阶段订购一件产品，以后各阶段应视具体情况而定．

参考文献

1 Bellman R. Dynamic programming. New Jersey: Princeton University Press, 1957.

2 Bellman R, Dreyfus S E. Applied dynamic Programming. New Jersey: Princeton University Press, 1962.

3 Dreyfus SE, Law A. The art and theory of dynamic programming. New York: Academic Press, 1977.

4 董加礼等编．工程运筹学．北京:北京工业大学出版社，1988.

5 董加礼．动态规划．长春:吉林工业大学应用数学系，1986.

6 刘光中．动态规划——理论及其算法．成都:成都科技大学应用数学系，1984.

第11篇

投入产出分析

编　者　秦学志　唐焕文
审校者　刘起运

目　录

引　言

投入产出分析这一学科的创始人瓦西里·列昂惕夫(Wassily Leontief),在其论著中,提出并阐述了投入产出模型和原理.**投入产出分析**是利用现代数学方法和计算机技术,来研究经济体系内各部门间投入与产出相互依存关系的数量经济分析方法.在美国和西欧常称之为投入产出技术、投入产出经济学或**投入产出法**等,在日本称之为**产业关联法**,前苏联等国家又称之为**部门联系平衡法**等.所谓**投入**,是指从事一项经济活动的消耗;所谓**产出**,是指从事一项经济活动的结果及其成果分配使用的去向.**投入产出模型**,从应用范围看,包括世界范围、几个国家或某个国家、几个地区或某个地区、部门、企业等;从计量单位看,可分为实物型的和价值型的等;从是否含时间变化的因素看,又可分为静态的和动态的,等等.

投入产出技术在国家经济工作中起着不可或缺的重要作用.它能较好地揭示国民经济中各种因素间的数量关系,为研究制订国民经济计划、中长期规划提供参考;可用来研究采取一项重要经济政策对整个经济结构的影响;可用来考察或确定产品的价格、计算工资或产品价格的变动对整个经济结构的影响;还可用来研究人口、就业、收入分配、国际贸易等各种社会经济问题,等等.

投入产出技术在世界范围内的发展方兴未艾,如把该技术与数学规划方法结合形成最优化模型,从静态模型向动态模型的发展与完善,利用计算机实现编表和建模的自动化,扩大研究和应用范围等.

最后需要指出:本篇各章的符号自成体系,同一符号在不同章中意义可能不同,特请注意.

1　全国静态产品投入产出模型

1.1　两种国民经济核算体系的投入产出表

1.1.1　国民经济核算体系

所谓**国民经济核算体系**,是指在一定经济理论指导下,综合运用统计学、会计学和数学等方法,为反映一个国家、地区在一定时期内经济活动成果所形成的社会核算指标系统.它由具有全面、科学和互相联系等特点的一系列指标组成,以此来较完整、准确地描述国民经济的结构和联系,反映国民经济生产、分配、交换和消费各领域和部门的经济运作状况.至今,世界各国采用的国民经济核算体系可分为两种:一种是 MPS 体系,即前苏联等国家所采用的"**国民经济平衡表体系**",亦称"物

质产品平衡表体系"或"东方核算体系";另一种是 SNA 体系,即实行市场经济的国家所采用的"国民经济账户体系",亦称"西方核算体系".目前,世界上采用 SNA 体系的国家较多.

1.1.2　MPS 核算体系与投入产出表

MPS 投入产出表是 MPS 核算体系的核心部分,主要反映物质生产部门的社会总产品、国民收入等指标情况,因此又称为物质产品投入产出表.自 1919 年,前苏联根据马克思、列宁关于社会再生产理论编制了"谷物饲料平衡表"开始,经多次修订和完善,1971 年由经互会统计委员会通过,联合国以经社部统计处名义发表了 MPS 的正式文件——《国民经济平衡表的基本原理》,逐步形成了 MPS 核算体系.简化的 MPS 投入产出表如表 1-1 所示.

表 1-1

<table>
<tr><td colspan="2" rowspan="4">投　入</td><td colspan="14">产　出</td><td rowspan="4">总产出</td></tr>
<tr><td colspan="5">中间产品</td><td colspan="9">最终产品</td></tr>
<tr><td rowspan="2">产品部门1</td><td rowspan="2">产品部门2</td><td rowspan="2">…</td><td rowspan="2">产品部门n</td><td rowspan="2">合计</td><td rowspan="2">固定资产更新、改造、大修理</td><td colspan="2">消费</td><td colspan="2">积累</td><td rowspan="2">进口(−)</td><td rowspan="2">出口(+)</td><td rowspan="2">合计</td></tr>
<tr><td>居民</td><td>社会</td><td>固定资产</td><td>流动资产</td></tr>
<tr><td rowspan="2">物质消耗</td><td>产品部门 1
产品部门 2
⋮
产品部门 n
合　计</td><td colspan="5">Ⅰ</td><td colspan="9" rowspan="2">Ⅱ</td><td rowspan="2"></td></tr>
<tr><td>固定资产折旧</td><td colspan="5"></td></tr>
<tr><td>活劳动消耗</td><td>劳动者报酬
社会纯收入
合　计</td><td colspan="5">Ⅲ</td><td colspan="9">Ⅳ</td><td></td></tr>
<tr><td colspan="2">总产值</td><td colspan="5"></td><td colspan="9"></td><td></td></tr>
</table>

1.1.3　SNA 核算体系的投入产出表

1968 年联合国发布了《国民经济核算体系》,即 SNA 新体系.在该体系中,所有账户都列在一个矩阵中,所有交易须发生在两个部门间或记录在一个账户的两个类目之间,新体系将投入产出表作为国民经济核算体系的重要组成部分.体系包含四个基本账户和十项交易,基本账户采用复式簿记形式,即每类都有它的借方和贷

方(或称为支出方和收入方).简化的 SNA 投入产出表如表 1-2 所示.

表 1-2

<table>
<tr><td colspan="3" rowspan="4">投　入</td><td colspan="7">产　出</td><td rowspan="4">总产出</td></tr>
<tr><td colspan="2">中间产品</td><td colspan="5">最终产品</td></tr>
<tr><td colspan="2">生　产</td><td colspan="2">消费
(消费品)</td><td colspan="2">积　累</td><td>国外</td></tr>
<tr><td>商品</td><td>产业部门</td><td>居民</td><td>政府</td><td>固定资产形成</td><td>储备增加</td><td>现期交易</td></tr>
<tr><td rowspan="2">中间投入</td><td rowspan="2">生产</td><td>商　品</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>产业部门</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">原始投入</td><td>消费
(收入)</td><td>固定资本折旧
雇员报酬
营业盈余
间接税净额</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>国外</td><td>现期交易</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="3">总投入</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

1.1.4 两种核算体系投入产出表的比较

1.理论基础不同

MPS 体系以马克思主义的经济学,特别是劳动价值论和社会再生产理论为基础;SNA 体系以西方近代的经济理论为基础,运用了瓦尔拉的一般均衡论.MPS 体系将投入产出表看做是一张产品平衡表,表中横行为实物分配使用情况,纵列为物质消耗构成,社会总产出仅限于物质产品和物质性劳务价值;SNA 体系将所有账户均列入一个矩阵中,将投入产出部门看做是生产账户的一个类目,借方为投入,贷方为产出,借贷平衡表现为总投入等于总产出,总产出包括市场出售的全部产品和劳务价值.MPS 体系将价值构成分为物化劳动的转移价值、社会必要劳动(为劳动者自己)创造的价值和剩余劳动(为社会)创造的价值;SNA 体系将总投入分为中间投入和原始投入(最初投入),原始投入包含的工资、租金、利息、利润被认为是劳动、土地、资本和企业家各生产要素所得的收入.

2.核算内容不同

因理论基础不同,MPS 核算体系与 SNA 核算体系在核算对象、方法、指标体系等方面都有差别.增加了劳务方面核算的新 MPS 体系,在核算内容上与 SNA 较接近.就产品流量核算而言,两种体系的主要差别在于对非物质服务流量的处理上.新 MPS 体系需另外设计一套非物质服务平衡表,详细地反映各非物质服务部门与国民经济其他部门的投入产出关系;SNA 体系直接把非物质服务流量放在产品流量核算中处理.就消费流量而言,MPS 体系把该项作为新创造价值,且仅限于物质

生产领域;SNA体系把消费收入作为原始投入,内容除全部生产要素的报酬外,还包括固定资产折旧.在消费支出核算中,MPS投入产出表的消费项目的内容仅限于个人和社会所消费的产品和物质性劳务;SNA投入产出表把它作为最终产品的组成部分,内容包括全部用于最终消费的产品和劳务,等等.

3.编表方法不同

MPS体系主要采用单式平衡表法,就其中某一表来看比较单一,只有综合使用这些平衡表才能较好地反映某些社会经济现象.SNA体系采用复式记账法和矩阵法.采用复式记账法,收支对应,易于编表和检验,可保证较高的编表质量;采用矩阵法,可根据需要将国民经济划分为若干个部门,使各部门收支关系系统化、条理化,由此可以清楚地掌握国民收入生产、分配、再分配和最终使用的全貌,有利于搞好国民经济综合平衡.矩阵式平衡表可以把几十、几百个账户归放在一起,表中横行代表收入,纵列代表支出,其系统完整地反映了各部门间的复杂联系,且只需从原矩阵中取出子矩阵即可,对账户所代表的经济部门进行更细致的划分和分析,并不需要改变原矩阵形式.

1.2　我国国民经济核算体系与投入产出表

我国先是学习并沿用前苏联50年代的统计模式,通过编制居民货币收支平衡表、物资平衡表、财政平衡表、试编的固定资产平衡表和部分人口平衡表等,建立国民经济核算体系.为适应新时期深化改革、加强宏观经济调控及与国际贸易接轨等的需要,在总结以往经验、吸收东西方核算体系优点的基础上,逐步建立起具有中国特色的国民经济核算体系及投入产出表.

1.2.1　中国特色的国民经济核算体系

1.我国国民经济核算体系的基本内容

(1)社会再生产条件的核算,包括人力、物力和财力的核算;

(2)社会再生产成果及其使用的核算,既进行物质产品生产及其使用的核算,又进行劳务活动及其使用的核算;

(3)社会再生产主要比例关系的核算;

(4)社会再生产效益的核算,等等.

2.方案表式的特征

方案基本表式分为六大部分,共16张表.为编制基本表式,还设计了第二层账户,账户按照再生产的投入、产出、流通、分配、消费与积累过程设置.方案既吸收了新MPS劳动平衡表的长处,又采纳了新SNA的优点,采取了与SNA相类似的结构,采用集合性和可分性原则,使之既可满足国内需要,又可用于国际对比,是一个体系较完善、逻辑较严密的系统.

1.2.2　中国特色的投入产出表

1987年,在具有中国特色的国民经济核算体系下形成了具有中国特色的投入

产出表,它既没有照搬 MPS 投入产出表,也没有照抄 SNA 投入产出表,而是从我国对内搞活、对外开放,大力推进社会、经济快速健康发展的需要出发,在吸收两大核算体系投入产出表长处的同时设计出来的,其表式如表 1-3 所示.

1.中国式投入产出表结构

中国式投入产出表由三个部分构成,分别以三个象限表示.第Ⅰ象限,宾栏是**中间使用**,主栏是**中间投入**,主、宾栏的各部门采用了双重分组,既有物质和非物质部门的分组,又有三个产业的分组.第Ⅰ象限主要反映国民经济各部门或三个产业之间的技术经济联系.第Ⅱ象限,宾栏是**最终使用**,主栏是中间投入,该象限主要反映社会产品和劳务用于消费、积累和出口等情况.第Ⅲ象限,宾栏是中间使用,主栏是**最初投入**,该象限主要反映的是国民经济各部门或三个产业在经济活动中需要消耗的固定资产、必要劳动和剩余劳动,表现为固定资产折旧、劳动者收入、福利基金、利润和税金等.第Ⅰ、Ⅱ象限合在一起反映的是全社会的物质产品和劳务的分配使用情况,第Ⅰ、Ⅲ象限合在一起反映的是国民经济各部门或三个产业的产品和劳务的价值构成.

2.中国式投入产出表的特点

(1)以马克思主义经济学原理为指导.它坚持了马克思主义关于物质生产部门与非物质生产部门的严格划分原则,分部门、分系统、分层次地核算国民经济这一有机整体及其内部相互制约的数量关系,从不同角度研究和揭示了社会经济的发展规模、结构和水平.

(2)适应于两种核算体系进行对比的需要.经过调整、转换,表 1-3 可分别与 MPS 或 SNA 核算体系的投入产出表进行对比.

(3)采用积木式板块结构的表式.板块结构具有易于拆卸、拼装、归并和转换等特点,可适应不同经济分析的需要.

1.3 全国静态产品投入产出模型

投入产出模型是投入产出理论的具体应用和表现形式.不含时间因素的投入产出模型称为**静态模型**.静态产品投入产出模型是静态模型的核心.其他各类静态模型,如地区投入产出模型、劳动力投入产出模型、固定资产投入产出模型等,可以看成是产品静态模型的扩充.

1.3.1 全国实物型静态产品投入产出模型

实物型投入产出模型是反映某一时期内(通常为 1 年)国民经济中用实物计量单位表示的各生产要素的投入使用状况或各部门产品流动分配状况的数学模型.投入产出模型是通过编制投入产出表而建立起来的,因此,首先讨论实物型投入产出表的表式结构,进而给出其模型.

1.实物型投入产出表

假设实物型投入产出表如表 1-4 所示.

表 1-3

（按当年生产

投入				产													
				中间使用													
				物质生产部门										非物质生产部门			
				第一产业			第二产业			第三产业							中间使用合计
				粮食作物种植业	…	第一产业合计	煤炭开采业	…	第二产业合计	铁路货运业	…	物质生产部门合计	铁路客运业	…	第三产业合计	非物质生产部门合计	
中间投入	物质生产部门	第一产业	粮食作物种植业 …														
			第一产业合计														
		第二产业	煤炭开采业 …														
			第二产业合计														
		第三产业	铁路货运业 …					I									
			物质生产部门合计														
	非物质生产部门		铁路客运业 …														
			第三产业合计														
			非物质生产部门合计														
中间投入合计																	
最初投入	固定资产折旧		基本折旧 大修理折旧														
			固定资产折旧合计														
	劳动者收入																
	其中：工资																
	福利基金																
	利润和税金											Ⅲ					
	其　他																
	最初投入合计																
总投入																	

者价格计算）　　　　单位:亿元

出																	
最终使用																进口	总产出
总消费					总积累									出口	最终使用合计		
居民总消费			社会总消费	总消费合计	固定资产形成					库存增加			总积累合计				
							基本建设										
农业居民	非农业居民	小计			固定资产大修理	固定资产更新改造	物质生产性	非物质生产性	小计	物质生产性	非物质生产性	小计					
							Ⅱ										

表 1-4

投入		产出						
		中间产品					最终产品	总产量
		部门 1	部门 2	…	部门 n	合计		
物质投入	部门 1	q_{11}	q_{12}	…	q_{1n}	$\sum_{j=1}^{n} q_{1j}$	Y_1	Q_1
	部门 2	q_{21}	q_{22}	…	q_{2n}	$\sum_{j=1}^{n} q_{2j}$	Y_2	Q_2
	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮
	部门 n	q_{n1}	q_{n2}	…	q_{nn}	$\sum_{j=1}^{n} q_{nj}$	Y_n	Q_n
劳动投入		q_{01}	q_{02}	…	q_{0n}	$\sum_{j=1}^{n} q_{0j}$	0	V

表中 q_{ij}表示j部门消耗i部门产品($i \neq 0$)或劳动投入($i=0$)的数量，$i=0,1,\cdots,n;j=1,2,\cdots,n$. $\sum_{j=1}^{n} q_{ij}$ 表示i部门($i \neq 0$)或劳动投入($i=0$)的中间产品总量，Y_i,Q_i 分别表示i部门($i \neq 0$)或劳动投入($i=0$)的最终产品量和总产出量. 从表的横行看，$q_{ij},j=1,2,\cdots,n,Y_i$ 合在一起表示i部门产品的分配使用状况，其中 $q_{ij},j=1,2,\cdots,n$ 表示i部门产品作为中间产品供各部门生产经营使用状况，Y_i 表示i部门产品作为最终产品供消费、积累、出口使用的状况，这两部分之和为在一定时期内i部门产品的生产总量，即产出总量 Q_i. 从表的纵列看，$q_{ij},i=0,1,\cdots,n$ 表示j部门为获得其产出总量需消耗所有部门产品的数量及劳动消耗量，其中劳动消耗量 q_{0j}可用日、时、货币等表示. 实物型投入产出表的同一横行各元素计量单位相同，同一纵列各元素的计量单位可能不同，因此，由表的横行可得到实物型投入产出表的基本方程.

2. 实物型投入产出表的基本方程

$$\sum_{j=1}^{n} q_{ij} + Y_i = Q_i, \quad i = 1,2,\cdots,n. \tag{1-1}$$

$$\sum_{j=1}^{n} q_{0j} = V. \tag{1-2}$$

3. 直接消耗系数

为进一步分析和预测，揭示部门间的生产技术经济联系和相互作用的内在规律，下面给出直接消耗系数的概念.

直接消耗系数表示某一产品的生产对其他产品的直接消耗程度，其表达式为

$$a_{ij} = q_{ij}/Q_j, \quad i,j = 1,2,\cdots,n. \tag{1-3}$$

确切地说，a_{ij}表示j部门生产一个单位的实物量需直接消耗i部门实物量的大小.

劳动的直接消耗系数为

$$a_{0j} = q_{0j}/Q_j, \quad j = 1,2,\cdots,n, \tag{1-4}$$

表示j部门生产一个单位的实物量需直接消耗劳动力数量的大小.

4. 实物型投入产出的基本数学模型

在一定时期内，若生产技术和中间产品的作用没有变化，则可以认为 $a_{ij}(i=0,1,\cdots,n;j=1,2,\cdots,n)$是相对稳定的，因此，由(1-3)式和(1-4)式可将(1-1)式和(1-2)式分别改写为

$$\sum_{j=1}^{n} a_{ij}Q_j + Y_i = Q_i, \quad i = 1,2,\cdots,n, \tag{1-5}$$

$$\sum_{j=1}^{n} a_{0j}Q_j = V. \tag{1-6}$$

记 $\boldsymbol{A}=(a_{ij})_{n\times n}$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$，$\boldsymbol{Q}=(Q_1,Q_2,\cdots,Q_n)^{\mathrm{T}}$，则(1-5)式可进一步写成

$$\boldsymbol{AQ}+\boldsymbol{Y}=\boldsymbol{Q}, \tag{1-7}$$

或

$$(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{Q}=\boldsymbol{Y}, \tag{1-8}$$

或

$$\boldsymbol{Q}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{Y}. \tag{1-9}$$

称 $\boldsymbol{A}$ 为直接消耗系数矩阵，$\boldsymbol{Q}$ 为各类产品的总产量列向量，$\boldsymbol{Y}$ 为各类产品的用于最终使用的列向量，$\boldsymbol{I}$ 为 n 阶单位阵.

(1-8)式和(1-9)式反映了最终产品和总产品之间的联系，已知总产品求最终产品可运用(1-8)式；反之，已知最终产品求总产品可运用(1-9)式.

1.3.2 全国价值型静态投入产出模型

价值型投入产出模型是反映某一时期内(通常为1年)国民经济中用统一货币计量单位表示的各生产要素投入使用状况或各部门产品流动分配状况的数学模型.价值型投入产出模型也是在价值型投入产出表的基础上建立起来的，假设价值型投入产出表如表1-5所示.

1. 价值型投入产出表

价值型投入产出表中的部门为**纯部门**，即每个部门都是同类产品(或劳务)的组合，如可将电冰箱、电视机、电风扇归入电气部门，并用统一货币单位表示其价值大小.表1-5的主栏由三部分组成，即物质消耗、活劳动消耗和总投入(总产值).**物质消耗**包括对劳动对象的消耗(即对各部门的消耗)和对固定资产的消耗(以折旧形式计入)两部分，**活劳动消耗**包括劳动报酬和社会纯收入两项，即**新创造价值**.编表时可将这两项展开，如可设计成工资、福利基金、劳动者收入、利润、税金、利息及其他等项.表1-5的宾栏包括中间产品、最终产品与总产品(总产出)三部分，与实

物型表类似.

表 1-5

<table>
<tr><td colspan="2" rowspan="3">投　　入</td><td colspan="10">产　　出</td></tr>
<tr><td colspan="5">中间产品</td><td colspan="4">最终产品</td><td rowspan="2">总产出</td></tr>
<tr><td>部门 1</td><td>部门 2</td><td>…</td><td>部门 n</td><td>小　计</td><td>固定资产更新、改造</td><td>消费</td><td>积累</td><td>小计</td></tr>
<tr><td rowspan="6">物质消耗</td><td>部门 1</td><td>x_{11}</td><td>x_{12}</td><td>…</td><td>x_{1n}</td><td>$\sum_{j=1}^{n} x_{1j}$</td><td>D_1</td><td>W_1</td><td>K_1</td><td>Y_1</td><td>X_1</td></tr>
<tr><td>部门 2</td><td>x_{21}</td><td>x_{22}</td><td>…</td><td>x_{2n}</td><td>$\sum_{j=1}^{n} x_{2j}$</td><td>D_2</td><td>W_2</td><td>K_2</td><td>Y_2</td><td>X_2</td></tr>
<tr><td>⋮</td><td>⋮</td><td>⋮</td><td></td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td></tr>
<tr><td>部门 n</td><td>x_{n1}</td><td>x_{n2}</td><td>…</td><td>x_{nn}</td><td>$\sum_{j=1}^{n} x_{nj}$</td><td>D_n</td><td>W_n</td><td>K_n</td><td>Y_n</td><td>X_n</td></tr>
<tr><td>小　计</td><td>$\sum_{i=1}^{n} x_{i1}$</td><td>$\sum_{i=1}^{n} x_{i2}$</td><td>…</td><td>$\sum_{i=1}^{n} x_{in}$</td><td>$\sum_{j=1}^{n}\sum_{i=1}^{n} x_{ij}$</td><td>$\sum_{i=1}^{n} D_i$</td><td>$\sum_{i=1}^{n} W_i$</td><td>$\sum_{i=1}^{n} K_i$</td><td>$\sum_{i=1}^{n} Y_i$</td><td>$\sum_{i=1}^{n} X_i$</td></tr>
<tr><td>折　旧</td><td>E_1</td><td>E_2</td><td>…</td><td>E_n</td><td>$\sum_{j=1}^{n} E_j$</td><td colspan="5"></td></tr>
<tr><td rowspan="3">活劳动消耗</td><td>劳动报酬</td><td>V_1</td><td>V_2</td><td>…</td><td>V_n</td><td>$\sum_{j=1}^{n} V_j$</td><td colspan="5"></td></tr>
<tr><td>社会纯收入</td><td>m_1</td><td>m_2</td><td>…</td><td>m_n</td><td>$\sum_{j=1}^{n} m_j$</td><td colspan="5"></td></tr>
<tr><td>小　计</td><td>N_1</td><td>N_2</td><td>…</td><td>N_n</td><td>$\sum_{j=1}^{n} N_j$</td><td colspan="5"></td></tr>
<tr><td colspan="2">总投入</td><td>X_1</td><td>X_2</td><td>…</td><td>X_n</td><td>$\sum_{j=1}^{n} X_j$</td><td colspan="5"></td></tr>
</table>

x_{ij}表示 j 部门生产价值量为 x_j时相应需要 i 部门投入的价值量,或需要消耗 i 部门的价值量,或表示 i 部门总产出 X_i 中投入 j 部门的价值量,故 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,n)$称为部门间的流出入量. Y_i 为 i 部门最终产品的产值, E_j, V_j, m_j 分别为 j 部门的折旧额、j 部门劳动者所得到的劳动报酬和 j 部门劳动者为社会创造的价值(社会纯收入), X_i 为 i 部门的总产值.

表 1-5 被分成四部分,习惯上按左上、右上、左下、右下的顺序分别称这四部分为Ⅰ、Ⅱ、Ⅲ、Ⅳ象限. 第Ⅰ象限是由 n 个部门纵横交叉组成的一张棋盘式表格,反映的是部门间的生产与分配的关系,为分析部门间的各种比例关系和运用数学方

法进行平衡计算提供了数据依据．折旧作为物质消耗，若放于第Ⅰ象限，则第Ⅰ象限就不是方阵，给数学处理，尤其给矩阵求逆带来一定不便，因此，通常将折旧单列出来成为Ⅰ、Ⅲ象限的中间项或放于第Ⅲ象限中．第Ⅱ象限表示各部门提供最终产品的数量，可体现为消费、积累和进出口等的比例及其构成，表 1-5 中未考虑进出口等因素．第Ⅲ象限主要反映各部门的净产出价值，体现了国民收入的初次分配及必要劳动和剩余劳动的比例及其构成，若折旧放在第Ⅲ象限，则第Ⅲ象限为各部门的增加值及其构成．第Ⅳ象限反映了某些国民收入的再分配情况，内容比较复杂，通常在编表时将该象限省略．

2. 价值型投入产出表的基本方程

由表 1-5 的横行可得

$$\sum_{j=1}^{n} x_{ij} + Y_i = X_i, \quad i = 1,2,\cdots,n. \tag{1-10}$$

由纵列可得

$$\sum_{i=1}^{n} x_{ij} + E_j + V_j + m_j = X_j, \quad j = 1,2,\cdots,n. \tag{1-11}$$

3. 价值型投入产出的基本数学模型

(1)直接消耗系数

记

$$a_{ij} = x_{ij}/X_j, \quad i,j = 1,2,\cdots,n. \tag{1-12}$$

称 $\boldsymbol{A} = (a_{ij})_{n\times n}$ 为直接消耗系数矩阵.

(2)基本数学模型

(1-10)式可改写为

$$\sum_{j=1}^{n} a_{ij}X_j + Y_i = X_i, \quad i = 1,2,\cdots,n, \tag{1-13}$$

或

$$\boldsymbol{AX} + \boldsymbol{Y} = \boldsymbol{X}, \quad \boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}, \tag{1-14}$$

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{Y}, \tag{1-15}$$

其中 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}, \boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^{\mathrm{T}}$.

(1-11)式可改写为

$$\sum_{i=1}^{n} a_{ij}X_j + E_j + V_j + m_j = X_j, \quad j = 1,2,\cdots,n, \tag{1-16}$$

记 $a_{C_j} = \sum_{i=1}^{n} a_{ij}, N_j = V_j + m_j$，则(1-16)式可写成

$$a_{C_j}X_j + E_j + N_j = X_j, \quad j = 1,2,\cdots,n, \tag{1-17}$$

或

$$(\boldsymbol{I} - \boldsymbol{A}_C)\boldsymbol{X} = \boldsymbol{E} + \boldsymbol{N}, \quad \boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{A}_C)^{-1}(\boldsymbol{E} + \boldsymbol{N}), \tag{1-18}$$

其中 I 为 n 阶单位阵，$\boldsymbol{A}_C = \mathrm{diag}(a_{C_1}, a_{C_2}, \cdots, a_{C_n})$，即 $\boldsymbol{A}_C$ 的主对角元素为 $a_{C1}, a_{C_2}, \cdots, a_{C_n}$，其他元素为 0，$\boldsymbol{E} = (E_1, E_2, \cdots, E_n)^{\mathrm{T}}, \boldsymbol{N} = (N_1, N_2, \cdots, N_n)^{\mathrm{T}}$.

定义 $a_{E_j}=E_j/X_j, a_{V_j}=V_j/X_j, a_{m_j}=m_j/X_j, a_{N_j}=N_j/X_j$ 分别为固定资产折旧系数、劳动报酬系数、社会纯收入系数和新创造价值系数，则有

$$a_{C_j}+a_{E_j}+a_{N_j}=1,\quad j=1,2,\cdots,n, \tag{1-19}$$

在不考虑进出口情况下，i 部门的生产量等于该部门的分配使用量，即

$$\sum_{j=1}^{n}x_{ij}+Y_i=\sum_{j=1}^{n}x_{ji}+E_i+N_i,\quad i=1,2,\cdots,n, \tag{1-20}$$

或

$$\sum_{j=1}^{n}a_{ij}X_j+Y_i=\sum_{j=1}^{n}a_{ji}X_i+E_i+N_i,\quad i=1,2,\cdots,n, \tag{1-21}$$

因此，

$$\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ij}+\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ji}+\sum_{i=1}^{n}E_i+\sum_{i=1}^{n}N_i, \tag{1-22}$$

或

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}X_j+\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ji}X_i+\sum_{i=1}^{n}E_i+\sum_{i=1}^{n}N_i, \tag{1-23}$$

从而

$$\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}(E_i+N_i), \tag{1-24}$$

即最终使用的合计值等于增加值的合计值.

1.4 完全消耗系数及其计算方法

产品在生产过程中除了与其他产品有直接联系外，还有间接联系，这种联系在各种产品的相互消耗中表现为，除有直接消耗外，还有间接消耗. 直接消耗和所有间接消耗之和即为完全消耗. 比如，建造一座大楼，需要直接消耗电力，同时又要直接消耗钢材、水泥、建筑设备等等，这些钢材、水泥、建筑设备等的生产又需要消耗电力，这样通过钢材等的生产形成了建造大楼对电力的一次间接消耗；而钢材等的生产同时需要消耗其他产品，这些其他产品的生产又需要消耗电力，形成了建造大楼对电力的二次间接消耗，依次还有三次、四次等间接消耗，因此建造大楼对电力的完全消耗即为其对电力的直接消耗和所有间接消耗之和.

完全消耗系数的计算公式为

$$\boldsymbol{B}=(b_{ij})_{n\times n}=\boldsymbol{A}+\boldsymbol{A}^2+\cdots+\boldsymbol{A}^m+\cdots, \tag{1-25}$$

或

$$\boldsymbol{B}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{A}=\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{A})^{-1}=(\boldsymbol{I}-\boldsymbol{A})^{-1}-\boldsymbol{I}, \tag{1-26}$$

其中 $\boldsymbol{A}$ 为直接消耗系数矩阵，$\boldsymbol{B}$ 为完全消耗系数矩阵，b_{ij}为 j 部门提供一个的单位最终产品对 i 部门的完全消耗系数.

下面以实例具体说明.

例 1 有 3 个部门，已知部门间的直接消耗系数矩阵为 $\boldsymbol{A}$，假定要求第 1 部门

最终提供1个单位的产品，试计算相应的直接消耗和间接消耗.其中

$$A=\begin{bmatrix}0.2 & 0.3 & 0.1\\ 0.1 & 0.2 & 0.4\\ 0.3 & 0.1 & 0.2\end{bmatrix}.$$

第1部门最终提供1个单位的产品，则对第1至第3部门的直接消耗为0.2,0.1,0.3,即为$A(1,0,0)^T$,记为$A^{(0)}$,而要使第1至第3部门直接提供产品$A^{(0)}$,又要消耗第1至第3部门的产品，其值为$A^{(1)}=AA^{(0)}=(0.1,0.16,0.13)^T$,即$A^{(1)}$为第1部门最终提供1个单位产品的一次间接消耗.同理可得二次间接消耗即$A^{(2)}=AA^{(1)}=(0.081,0.094,0.072)^T,\cdots$,则第1部门最终提供1个单位的产品需要的完全消耗为$A^{(0)}+A^{(1)}+\cdots+A^{(m)}+\cdots$.一般而言，间接消耗将随着次数的增大而逐次减小，以至微不足道，因此完全消耗量可以求得.

要求第2部门或第3部门最终提供一个单位产品时，相应的完全消耗量也可类似求得.综上分析便可给出(1-25)式或(1-26)式.

1.5 全国静态产品投入产出模型的应用

投入产出表包含十分丰富的数据资料，具有特殊的结构形式和运算方式，有着多种多样的应用.这里仅介绍它在经济分析、计划工作和价格分析中的应用.

1.5.1 在经济分析中的应用

经济分析的目的在于总结经济工作的成败经验与教训，分析国家、地区、部门或企业的优势或劣势，揭示国民经济各部门与社会再生产各环节间的内在联系，有的放矢地提高经济工作水平，使国民经济协调稳步向前发展.

1.分析国民经济中的一些基本比例关系

要使国民经济协调稳步发展，必须协调国民经济中各种主要比例关系，如两大部类的比例，农、轻、重的比例，积累与消费的比例等.投入产出模型能从再生产的角度提供深入研究与分析这些比例关系的数据资料，从具体的数量关系上揭示这些比例关系的实质.下面以表1-6(价值型投入产出表)为例说明.

由表1-6可得到相应的直接消耗系数和完全消耗系数，并分别列于表1-7和表1-8.

(1)分析两大部类比例关系.**两大部类产品**在生产与分配使用之间保持一定的比例关系，是使社会再生产得以顺利进行的重要条件之一，即两大部类产品不仅在实物形态上要顺利地实现交换，而且在价值形态上也要能得到补偿.利用价值型产品投入产出表可以较精确地计算出在社会产品中两大部类产品的具体数量和价值构成.其计算过程分别为

1)计算生产资料与消费资料总量.根据按产品实际用途划分两大部类的原则，生产资料包括生产中消耗的劳动对象(即中间产品)和用于扩大再生产的组成生产性固定资产积累与流动资产积累的产品；消费资料包括用于本期及今后时期个人与社会消费的产品，含新增的非生产性固定资产(如住宅、文化福利设施等)和消费

表 1-6　　单位:亿元

投入		产出									
		中间产品					最终产品				总产出
		农业	轻工业	重工业	其他	小计	消费(W_i)	生产性积累(K_i)	非生产性积累(F_i)	小计	
物质消耗	农业	108	230	182	170	690	820	190	100	1110	1800
	轻工业	18	690	78	170	956	1100	160	84	1344	2300
	重工业	360	345	1040	340	2085	170	250	95	515	2600
	其他	54	345	260	170	829	590	161	120	871	1700
	小计	540	1610	1560	850	4560	2680	761	399	3840	8400
新创造价值	劳动报酬(V_j)	1070	240	490	400	2200					
	社会纯收入(m_j)	190	450	550	450	1640					
	小计(N_j)	1260	690	1040	850	3840					
总产值(X_j)		1800	2300	2600	1700	8400					

表 1-7

	农业	轻工业	重工业	其他
农业	0.06	0.10	0.07	0.10
轻工业	0.01	0.30	0.03	0.10
重工业	0.20	0.15	0.40	0.20
其他	0.03	0.15	0.10	0.10

表 1-8

	农业	轻工业	重工业	其他
农业	0.1090	0.2356	0.1725	0.1877
轻工业	0.0464	0.5018	0.1134	0.1972
重工业	0.4114	0.5608	0.8284	0.5143
其他	0.0904	0.3205	0.2278	0.2074

品的库存和储备增加额等．在表 1-6 中第一部类产品为(4560 + 761)亿元 = 5321 亿元，第二部类产品为(2680 + 399)亿元 = 3079 亿元．

2)计算各种消耗系数．各部门的物质消耗系数 a_{C_j}，劳动报酬系数 a_{V_j} 和社会纯收入系数 a_{m_j} 分别为

$$a_{C_j} = \sum_{i=1}^{4} a_{ij},\ a_{V_j} = V_j/X_j,\ a_{m_j} = m_j/X_j,\ j = 1,2,\cdots,4,$$

其中 $A = (a_{ij})_{4\times4}$ 为直接消耗系数矩阵．

计算结果如表 1-9 所示．

表 1-9

	农业	轻工业	重工业	其他
a_{C_j}	0.300	0.700	0.600	0.500
a_{V_j}	0.594	0.104	0.188	0.235
a_{m_j}	0.106	0.196	0.212	0.265
合计	1.000	1.000	1.000	1.000

3)计算第二部类的价值构成．物质消耗，即第二部类的转移价值．

$$C_{\text{II}} = \sum_{j=1}^{4} a_{C_j}(W_j + F_j), \tag{1-27}$$

劳动报酬
$$V_{\text{II}} = \sum_{j=1}^{4} a_{V_j}(W_j + F_j), \tag{1-28}$$

社会纯收入
$$m_{\text{II}} = \sum_{j=1}^{4} a_{m_j}(W_j + F_j), \tag{1-29}$$

其中 W_j 和 $F_j(j=1,2,\cdots,4)$ 的意义见表 1-6．

由表 1-6 和表 1-9 可得

$C_{\text{II}} = 0.3\times(820+100)$亿元 $+0.7\times(1100+84)$亿元 $+0.6\times(170+95)$亿元 $+$
$0.5\times(590+120)$亿元 $=1618.80$ 亿元，

$V_{\text{II}} = 0.594\times(820+100)$亿元 $+0.104\times(1100+84)$亿元 $+0.188\times(170+95)$亿元 $+$
$0.235\times(590+120)$亿元 $=886.29$ 亿元，

$m_{\text{II}} = 0.106\times(820+100)$亿元 $+0.196\times(1100+84)$亿元 $+0.212\times(170+95)$亿元 $+$
$0.265\times(590+120)$亿元 $=573.91$ 亿元．

4)计算第一部类的价值构成．

物质消耗
$$C_{\text{I}} = \sum_{i=1}^{4}\sum_{j=1}^{4} x_{ij} - C_{\text{II}}, \tag{1-30}$$

劳动报酬
$$V_{\text{I}} = \sum_{j=1}^{4} V_j - V_{\text{II}}, \tag{1-31}$$

社会纯收入　　　$m_{\mathrm{I}} = \sum_{j=1}^{4} m_j - m_{\mathrm{II}}$,　　　(1-32)

由表 1-6 和 3)中计算结果可得

$$C_{\mathrm{I}} = (4560 - 1618.80)\text{亿元} = 2941.20\text{ 亿元},$$
$$V_{\mathrm{I}} = (2200 - 886.29)\text{亿元} = 1313.71\text{ 亿元},$$
$$m_{\mathrm{I}} = (1640 - 573.91)\text{亿元} = 1066.09\text{ 亿元},$$

将上述结果及某些比例值列于表 1-10 中.

表 1-10　　　　单位:亿元

	总产值	物质消耗(C)		劳动报酬(V)		社会纯收入(m)		国民收入	
		数量	占总产值(%)	数量	占总产值(%)	数量	占总产值(%)	数量	占总产值(%)
合计	8400	4560	54.3	2200	26.2	1640	19.5	3840	45.7
第一部类	5321	2941.20	55.3	1313.71	24.7	1066.09	20.0	2379.8	44.7
第一部类占总产值/(%)	63.3	64.6	—	59.7	—	65.0	—	61.9	—
第二部类	3079	1618.80	52.5	886.29	28.8	573.91	18.7	1460.2	47.5
第二部类占总产值/(%)	36.7	35.4	—	40.3	—	35.0	—	38.1	—

由表 1-10 可见,第一部类产品为 5321 亿元,占全部社会产品的 63.3%,第二部类产品为 3079 亿元,占全部社会产品的 36.7%.由$(C+V+m)_{\mathrm{I}} - (C_{\mathrm{I}} + C_{\mathrm{II}}) = V_{\mathrm{I}} + m_{\mathrm{I}} - C_{\mathrm{II}} = (2379.8 - 1618.8)$亿元$= 761$ 亿元,此值即为该时期可用来扩大再生产的生产资料数量,为最终产品中用于生产性积累的数额,表明了有占全部社会产品的 9.06% 的生产资料可用于扩大再生产.

欲使扩大再生产顺利进行,需要有消费资料的积累,即须满足

$$(C+V+m)_{\mathrm{II}} > V_{\mathrm{I}} + \left(\frac{m}{x}\right)_{\mathrm{I}} + V_{\mathrm{II}} + \left(\frac{m}{x}\right)_{\mathrm{II}},$$

其中 m/x 在资本主义社会指资本家的消费部分,在社会主义社会可理解为非生产部门人员的消费.公式表明,第二部类的产品应大于两大部类工人和资本家(非生产部门人员)的消费.由表 1-10 知,$(C+V+m)_{\mathrm{II}} = 3079$ 亿元,由表 1-6 知,$V_{\mathrm{I}} + V_{\mathrm{II}} + (m/x)_{\mathrm{I}} + (m/x)_{\mathrm{II}} = \sum_{i=1}^{4} W_i = 2680$亿元,所以,$(C+V+m)_{\mathrm{II}} - (V_{\mathrm{I}} + V_{\mathrm{II}} + (m/x)_{\mathrm{I}} + (m/x)_{\mathrm{II}}) = 399$亿元$= \sum_{i=1}^{4} F_i$,表明第二部类产品除满足本期社会消

费外，还剩余 399 亿元作为消费积累，这为进一步扩大再生产创造了条件.

(2)分析农、轻、重比例关系. 农业、轻工业和重工业的比例关系协调与否，从根本上决定着整个国民经济能否健康稳定地向前发展，是国民经济的一项重要的比例关系. 利用投入产出模型可以分析农、轻、重的内部结构，了解它们各自的具体部门构成和价值构成，从社会再生产的角度研究分析它们之间的内在联系等.

1)分配系数、最终产品结构系数. 在抽象了进出口因素的条件下，**分配系数**可定义为

$$h_{ij}=\begin{cases}x_{ij}/X_i, & j=1,2,\cdots,n;\\ Y_{ip}/X_i, & j=n+p, p=1,2,\cdots,l.\end{cases} \tag{1-33}$$

其中 x_{ij}为 i 部门的中间产品，Y_{ip}为 i 部门产品用于第 p 项最终产品的数量，假设最终产品有 l 项.

分配系数反映的是各部门产品的分配使用构成及其比例.

最终产品结构系数定义为

$$d_{ip}=Y_{ip}/U_p, \quad i=1,2,\cdots,n; p=1,2,\cdots,l, \tag{1-34}$$

其中 U_p 为第 p 项最终产品的总量.

最终产品结构系数反映的是最终产品各项的构成及其比例.

仍以表 1-6 为例，相应的分配系数和最终产品结构系数分别如表 1-11 和表 1-12 所示.

表 1-11　　单位：%

	中间产品					最终产品			各部门产品占总产品比例
	农业	轻工业	重工业	其他	小计	消费	积累	小计	
农业	6.0	12.8	10.0	9.4	38.3	45.6	16.1	61.7	21.4
轻工业	0.8	30.0	3.4	7.4	41.6	47.8	10.6	58.4	27.4
重工业	13.8	13.3	40.0	13.1	80.2	6.5	13.3	19.8	30.9
其他	3.2	20.3	15.3	10.0	48.8	34.7	16.5	51.2	20.3
社会总产品分配比例	6.4	19.2	18.6	10.1	54.3	31.9	13.8	45.7	100.0

表 1-12　　单位：%

	消　费	积　累		小　计
		生产性	非生产性	
农业	30.6	25.0	25.0	28.9
轻工业	41.1	21.0	21.0	35.0
重工业	6.3	32.9	23.8	13.4
其他	22.0	21.1	30.2	22.7
小计	100.0	100.0	100.0	100.0

表 1-11 给出了农、轻、重各部门产品的分配使用情况，即各部门产品用于生产消耗、消费和积累的比重及各部门产品占总产品的比重，这些数据为计划工作提供了重要的参考资料. 配合表 1-7 和表 1-8 所提供的直接消耗系数和完全消耗系数，可以了解农、轻、重各部门的内在联系. 表 1-12 给出了最终产品各项中由各部门提供的比例，由此可了解最终产品各项的来源及其构成.

2)结合分析农、轻、重的比重与两大部类的比重. 通过计算农业、轻工业和重工业产品中，生产资料与消费资料的数量及比重，可了解各部门产品满足生产发展需要与满足人民生活需要的情况.

以表 1-6 为例计算各部门两大部类产品数量及比重，结果如表 1-13 所示.

表 1-13

		第一部类产品	第二部类产品	总　计
农业	数量/亿元	880	920	1800
	占总计/(%)	48.9	51.1	—
轻工业	数量/亿元	1116	1184	2300
	占总计/(%)	48.5	51.5	—
重工业	数量/亿元	2335	265	2600
	占总计/(%)	89.8	10.2	—
其他	数量/亿元	990	710	1700
	占总计/(%)	58.2	41.8	—

3)分析农、轻、重比例是否协调. 依据再生产原理，从生产能否满足需要出发衡量农、轻、重比例是否协调. 因农业、轻工业产品大部分直接用于人民消费的需要，重工业产品主要用于满足发展生产的需要，所以，可采用下列准则判断农、轻、重比例是否协调：农业、轻工业产品在扣除用于生产的消耗后，其剩余部分能否满足人民生活水平提高的需要；重工业产品在扣除用于生产的消耗后，其剩余部分能否满足发展生产的需要.

(3)分析积累与消费的比例. 积累与消费的比例关系是关系到发展生产、改善人民生活水平的重要比例关系，在一定程度上反映了国家、集体、个人三者利益关系是否协调合理. 在分析这个比例时，不仅要看积累率的高低，而且要结合各时期的生产发展情况，具体衡量积累与消费的安排能否适应国民经济发展的需要，适应实际生产可能提供的条件. 表 1-11 和表 1-12 分别给出了消费和积累在各部门总产品中的比重及在消费和积累的总额中各部门所占的比重，这些分别说明了各部门每一个单位产品中消费和积累所占的份额及国民经济中每一个单位的消费或积累由各部门提供的份额.

下面讨论积累与消费总量间的比例关系. 该比例关系以国民收入使用额为总

体，计算积累额和消费额所占的比重，即给出积累率和消费率. 国民收入使用额在实物形态上指扣除国家储备、进出口差额和固定资产更新、大修理后剩余的最终产品数量，在价值形态上为净产值，即国民收入. 由此，积累率可定义为

$$a_K = \sum_{i=1}^{n} K_i \Big/ \sum_{i=1}^{n} Y_i, \quad 或 \quad a_K = \sum_{i=1}^{n} K_i \Big/ \sum_{i=1}^{n} (V_i + m_i), \tag{1-35}$$

其中 $\sum_{i=1}^{n} K_i$ 为各部门提供的积累总额，$\sum_{i=1}^{n} Y_i$ 为最终产品量，$\sum_{i=1}^{n}(V_i + m_i)$ 为国民收入总额.

消费率定义为

$$a_W = 1 - a_K. \tag{1-36}$$

另外，还可以以国民收入为总体，分析积累与消费的关系，或者单以消费或积累为总体分析其各种构成的关系等.

(4)分析国民经济各部门间的比例关系. 国民经济作为一个有机整体，各部门间存在着极其复杂的技术经济联系，弄清这些联系可以为建立合理的产业结构提供科学的依据.

1)直接消耗系数和完全消耗系数的分析应用. 直接消耗系数和完全消耗系数可以从数量关系上揭示各部门间内在的直接和间接的技术经济联系. 直接消耗系数矩阵中每列的数字反映了该列所对应部门对所有部门的直接依赖程度，完全消耗系数矩阵中每列的数字反映了该列所对应部门对所有部门的直接和间接依赖程度.

2)感应度系数与影响力系数的分析应用.

影响力系数 π_j 定义为

$$\pi_j = \sum_{i=1}^{n} b_{ij}^* \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^*\right), \quad j = 1,2,\cdots,n, \tag{1-37}$$

其中 $B^* = (b_{ij}^*)_{n\times n} = (I - A)^{-1}$，$A$ 为直接消耗系数矩阵，I 为 n 阶单位阵.

感应度系数 σ_i 定义为①

$$\sigma_i = \sum_{j=1}^{n} h_{ij}^* \Big/ \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^*\right), \quad i = 1,2,\cdots,n, \tag{1-38}$$

其中 $H^* = (h_{ij}^*)_{n\times n} = (I - H)^{-1}$，$H = (h_{ij})_{n\times n}$，$h_{ij} = x_{ij}/X_i$，$x_{ij}$，$X_i$ 的意义同前.

影响力系数亦称为后向联系系数，π_j 表示 j 部门最终产品每增加一个单位时，需国民经济各部门增加生产的产品总量，与国民经济中各部门均分别增加一个单位最终产品时平均使各部门增加生产产品总量的比较. π_j 值有三种可能：$\pi_j > 1$，$\pi_j = 1$ 和 $\pi_j < 1$，其中 $\pi_j > 1$ 表明 j 部门生产的发展需国民经济各部门较快速度发展来相配合，是提高国民经济发展速度的关键部门. 感应度系数亦称为前向联系系数，表明 i 部门增加单位增加值对各部门产出的推动程度.

① 陈锡康. 投入占用产出的理论及其应用. 见：李强，刘起运主编，当代中国投入产出实践与研究. 北京：中国统计出版社，1999.

3)计算各部门物质消耗比重.

$$a_{ij}^{c} = x_{ij} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}, \quad i,j = 1,2,\cdots,n, \tag{1-39}$$

a_{ij}^{c}表示j部门消耗i部门的产品在整个中间消耗中所占的比重,$(a_{ij}^{c})_{n\times n}$在一定程度上反映了各部门在生产过程中的相互依赖程度.

4)计算各部门新创造价值在国民收入中所占的比重.

j部门新创造价值占国民收入比重的计算公式为

$$V_j \Big/ \sum_{i=1}^{n}(V_i + m_i) \text{ 和 } m_j \Big/ \sum_{i=1}^{n}(V_i + m_i) \text{ 或}(V_j + m_j) \Big/ \sum_{i=1}^{n}(V_i + m_i). \tag{1-40}$$

2.分析国民经济的宏观经济效果

可从宏观和微观两个角度考察经济效果.**宏观经济效果**指整个国民经济范围内全社会的经济运作效果,微观经济效果指相对较小范围内的经济运作效果.考察经济效果一般可用下列指标.

(1)物化劳动消耗效果.指物化劳动消耗与有用成果间的比较.物化劳动消耗指各部门或整个国民经济的物质消耗总量,有用成果指各部门的总产出或社会总产出,或者指各部门的净产品或社会净产品,或者指各部门的纯收入或社会纯收入.等等.与之相应,可计算出各部门(或社会)物耗总产出率、各部门(或社会)净产值率、各部门(或社会)物耗纯收入率等.例如,社会物耗净产值率为

$$(V + m)/C = \sum_{j=1}^{n}(V_j + m_j) \Big/ \left(\sum_{j=1}^{n}\sum_{i=1}^{n} x_{ij} + \sum_{j=1}^{n} E_j\right), \tag{1-41}$$

j部门物耗净产值率为

$$(V_j + m_j) \Big/ \left(\sum_{i=1}^{n} x_{ij} + E_j\right), \tag{1-42}$$

其中 E_j 为j部门固定资产折旧,其他符号意义同前.

(2)活劳动消耗效果.指活劳动消耗量与有用成果间的比较.活劳动消耗量指各部门或整个国民经济的劳动报酬总量.相应地可分别得到劳动报酬总产出率、劳动报酬净产值率和劳动报酬利税率等.即

劳动报酬总产出率 $\overline{X}/V$(或 x_j/V_j), (1-43)

劳动报酬净产值率 $(V+M)/V$(或$(V_j+M_j)/V_j$), (1-44)

劳动报酬利税率 M/V(或 m_j/V_j), (1-45)

其中 $\overline{X} = \sum_{i=1}^{n} x_i, V = \sum_{j=1}^{n} V_j, M = \sum_{j=1}^{n} m_j$.

(3)物化劳动和活劳动的综合效果.是指生产成本与有用成果间的比较.相应地可分别得到成本总产出率、成本净产值率、成本纯收入率等.即

成本总产出率 $\overline{X}/(C+V)$(或 $x_j/(C_j+V_j)$), (1-46)

成本净产值率 $(V+M)/(C+V)$(或$(V_j+m_j)/(C_j+V_j)$), (1-47)

成本纯收入率 $M/(C+V)$(或 $M_j/(C_j+V_j)$), (1-48)

(4)劳动消耗获得的消费资料的效果.指劳动消耗与获得的消费资料的比较.相应地有物耗消费资料率、劳动报酬消费资料率、生产成本消费资料率等.即

物耗消费资料率　　W/C(或 W_j/C_j), (1-49)

劳动报酬消费资料率　　W/V(或 W_j/V_j), (1-50)

成本消费资料率　　$W/(C+V)$(或 $W_j/(C_j+V_j)$), (1-51)

其中,$W = \sum_{j=1}^{n} W_j$.

上述各项指标,其值越大表明其对应的部门或社会经济效果越好;反之,其值越小,其对应的部门或社会经济效果越差.一般而言,经济效果好的部门可以考虑加快其发展速度,但由于各部门间存在着复杂的技术经济联系,因而一部门可否加快发展速度,还取决于与它相联系的其他部门的发展状况.

1.5.2　在价格分析中的应用

一个合理的价格体系既能对生产、分配、交换、消费各环节起调节作用,又有利于国民经济的协调稳步发展.合理的价格体系的制定是一个复杂的问题,不仅需要作理论上的探讨,还需要有一套科学的计算方法.投入产出表提供的资料有助于对价格的形成与各种产品价格间的相互影响进行分析.

1.价格模型

商品价格是由生产该商品的社会必要劳动量决定的,是商品价值的货币表现.价值型投入产出表反映了各部门产品的价值构成,即转移价值(C)、劳动者为自己创造的价值(V)和劳动者为社会创造的价值(m).若各部门劳动者的劳动报酬能够依据"按劳分配"的原则正确确定,各部门(各种产品)的盈利率也可按照统一的标准确定,则可利用投入产出模型计算各部门产品的价格.

记 $q_{ij}, Y_i, Q_i, a_{ij}, i=1,2,\cdots,n; j=1,2,\cdots,n$ 分别为以实物计量单位表示的部门间的流出入量、部门最终产品、总产品和直接消耗系数;V_j, m_j, E_j 和 P_j 分别为以货币单位计量的 j 部门的劳动报酬、社会纯收入、固定资产折旧和产品单价,则由价值型投入产出模型,即(1-11)式,有

$$\sum_{i=1}^{n} q_{ij}P_i + E_j + V_j + m_j = Q_jP_j,\quad j = 1,2,\cdots,n,$$

或

$$\sum_{i=1}^{n} a_{ij}P_i + a_{E_j} + a_{v_j} + a_{m_j} = P_j,\quad j = 1,2,\cdots,n, \tag{1-52}$$

其中 $a_{E_j} = E_j/Q_j, a_{V_j} = V_j/Q_j, a_{m_j} = m_j/Q_j, a_{ij} = P_{ij}/Q_j$.

记 $\overline{E} = (a_{E_1}, a_{E_2}, \cdots, a_{E_n})^{\mathrm{T}}, \overline{V} = (a_{V_1}, a_{V_2}, \cdots, a_{V_n})^{\mathrm{T}}, \overline{M} = (a_{m_1}, a_{m_2}, \cdots, a_{m_n})^{\mathrm{T}}, P = (P_1, P_2, \cdots, P_n)^{\mathrm{T}}, A = (a_{ij})_{n\times n}$,则(1-52)式可写为

或

$$A^{\mathrm{T}}P + \overline{E} + \overline{V} + \overline{M} = P \quad 或 \quad P = (I - A^{\mathrm{T}})^{-1}(\overline{E} + \overline{V} + \overline{M}),$$

$$P = [(I-A)^{-1}]^{T}(\overline{E}+\overline{V}+\overline{M}), \tag{1-53}$$

(1-53)式即为价格模型.由此可得价格变动模型

$$\Delta P = [(I-A)^{-1}]^{T}(\Delta\overline{E}+\Delta\overline{V}+\Delta\overline{M}), \tag{1-54}$$

它揭示了 $\overline{E}$,$\overline{V}$ 或 $\overline{M}$ 的变化对价格变化的影响.

2. 价格影响模型

价格影响模型是分析某部门(或某些部门)产品价格变动对其他部门产品价格影响的数学模型.某一部门(或某些部门)的价格变动,会使与其有联系的部门产品成本发生变化,从而引起产品价格的变化.假设第 n 部门的价格变动为 ΔP_n,则由此引起的其他部门的价格变动满足

$$(\Delta P_1,\Delta P_2,\cdots,\Delta P_{n-1})^{T} = A_{n-1}^{T}(\Delta P_1,\Delta P_2,\cdots,\Delta P_{n-1})^{T} + (a_{n1},a_{n2},\cdots,a_{n,n-1})^{T}\Delta P_n,$$

或

$$\begin{aligned}(\Delta P_1,\Delta P_2,\cdots,\Delta P_{n-1})^{T} &= (I-A_{n-1}^{T})^{-1}(a_{n1},a_{n2},\cdots,a_{n,n-1})^{T}\Delta P_n \\ &= [(I-A_{n-1})^{-1}]^{T}\cdot(a_{n1},a_{n2},\cdots,a_{n,n-1})^{T}\Delta P_n,\end{aligned} \tag{1-55}$$

其中

$$A_{n-1} = \begin{bmatrix} a_{11} & \cdots & a_{1,n-1} \\ \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n-1} \end{bmatrix}.$$

1.5.3 在计划工作中的应用

编制国民经济计划的目的,是在遵循经济规律的条件下,正确确定国民经济中各种比例关系,使整个国民经济高效、协调、稳步地向前发展,同时使社会生产不断满足人民生活水平提高的需要.投入产出表能提供许多重要参数,使计划工作既能从局部与全局相结合的角度建立社会生产、分配、交换、消费各环节的平衡,又能做到社会产品在实物形态与价值形态方面的平衡,还可以从再生产的角度考察积累与消费的影响等,因此投入产出模型在编制国民经济计划工作中起着重要的作用.

1. 为编制国民经济计划提供了一种科学方法

社会生产的协调、高效、稳步发展,需要保持国民经济中各种比例关系适宜,需要社会产品在各个环节中达到平衡,因此盲目追求社会生产高速发展和高消费,都是有悖于经济发展规律的.投入产出表提供了一个从最终产品出发确定国民经济各部门总产量的数学模型,在一定程度上解决了从满足社会需要出发制定国民经济计划的具体方法问题,配合采用在计划时期要达到的目标与现实经济所能提供的条件之间进行综合平衡等手段,可以编制出较理想的经济计划.

(1)从最终产品出发编制国民经济计划.投入产出表中的最终产品,其主要项目是消费和积累,所以利用模型从最终产品出发来编制经济计划,基本上可体现扩大再生产和满足人民生活消费需要的目的.具体过程可概括为

1)确定计划期应达到的消费总额及其构成.主要根据人口的变化状况、人民生活需要提高的程度及计划期人民消费的总需求量等因素确定.参照报告期的最终产品结构系数表,考虑人口的变化、新增的劳动力及其他因素对消费需求的影响,

可科学地预测居民、社会消费构成的变化情况，确定出计划期应达到的消费总额及各部门可预期提供的消费额.

2)确定计划期应达到的积累总额及其构成. 可在先确定计划期国民收入的基础上，由计划期的积累率确定积累总额，也可用报告期固定资产投入产出表中的固定资金直接占用系数和完全占用系数，并考虑计划期最终产品的增长情况来确定固定资产的积累额. 计划期内流动资金的积累额也可类似确定. 参照报告期的最终产品结构系数表，经预测调整可给出计划期积累的部门构成.

3)确定计划期的直接消耗系数. 在短期计划中，一般可在报告期直接消耗系数的基础上，对个别采用了新技术或变化较大的部门所对应的系数进行修正. 对于长期计划，需进行局部或全面的修正，具体可参照的方法，如 RAS 法等将在本篇第 2 章介绍.

4)确定计划期各部门的总产出. 由 $\boldsymbol{X}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{Y}$ 计算出计划期各部门的总产出，即由社会最终产品量确定的各部门必须达到的产量，因此最终产品中消费额和积累额确定得是否合适，还需将由此计算得到的各部门必须达到的产量同各部门的实际能力进行比较、平衡，须经反复调整，最终才能确定出适宜的计划期内各部门的最终产品量和产量.

(2)从最终产品与总产品相结合出发，编制国民经济计划. 单从最终产品出发编制国民经济计划，虽然模型简单，但也存在不足：一是某些部门中最终产品的数据难以预测和确定，二是在反复调整的过程中需参照各部门的产量. 因此，可以说从最终产品出发编制国民经济计划的方法实质上亦为一种从最终产品和总产品相结合出发编制国民经济计划的方法. 为了区别，此处从最终产品与总产品相结合出发编制国民经济计划，指的是某些部门最终产品中的数据无法预测和确定的情况，具体做法如下.

1)确定最终产品和总产品的已知量和未知量.

已知量指经预测可以确定的量，不易确定的量作为未知量.

2)建立数学模型.

$$\boldsymbol{X}^{(t)}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{Y}^{(t)},\tag{1-56}$$

其中，

$$\boldsymbol{C}=(\boldsymbol{I}-\boldsymbol{A})^{-1}=\begin{bmatrix}\boldsymbol{C}_{m,m} & \boldsymbol{C}_{m,n-m}\\ \boldsymbol{C}_{n-m,m} & \boldsymbol{C}_{n-m,n-m}\end{bmatrix};$$

$\boldsymbol{Y}^{(t)}=((\boldsymbol{Y}_m^1)^{\mathrm{T}},(\boldsymbol{Y}_{n-m}^2)^{\mathrm{T}})^{\mathrm{T}}$，$\boldsymbol{X}^{(t)}=((\boldsymbol{X}_m^1)^{\mathrm{T}},(\boldsymbol{X}_{n-m}^2)^{\mathrm{T}})^{\mathrm{T}}$，均为列向量；$\boldsymbol{X}_m^1$，$\boldsymbol{Y}_{n-m}^2$分别为由计划期社会总产品中 m 种产品的已知量组成的列向量和 $n-m$ 种最终产品的已知量组成的列向量；$\boldsymbol{X}_{n-m}^2$，$\boldsymbol{Y}_m^1$ 分别为由计划期社会总产品中 $n-m$ 种产品的未知量组成的列向量和 m 种最终产品的未知量组成的列向量.

(1-56)式可展开成

$$\boldsymbol{X}_m^1=\boldsymbol{C}_{m,m}\boldsymbol{Y}_m^1+\boldsymbol{C}_{m,n-m}\boldsymbol{Y}_{n-m}^2,\tag{1-57}$$

$$\boldsymbol{X}_{n-m}^2=\boldsymbol{C}_{n-m,m}\boldsymbol{Y}_m^1+\boldsymbol{C}_{n-m,n-m}\boldsymbol{Y}_{n-m}^2.\tag{1-58}$$

由(1-57)，(1-58)式可得

$$\boldsymbol{Y}_m^1=\boldsymbol{C}_{m,m}^{-1}(\boldsymbol{X}_m^1-\boldsymbol{C}_{m,n-m}\boldsymbol{Y}_{n-m}^2),\tag{1-59}$$

$$X_{n-m}^2 = C_{n-m,m}C_{m,m}^{-1}(X_m^1 - C_{m,n-m}Y_{n-m}^2) + C_{n-m,n-m}Y_{n-m}^2. \qquad (1\text{-}60)$$

2. 投入产出模型在计划调整方面的应用

计划在实施过程中不是一成不变的,需要根据实际情况进行调整,调整时可采用公式 $\Delta X = (I-A)^{-1}\Delta Y$,或 $\Delta Y = (I-A)\Delta X$ 计算最终产品的变化对各部门产量需求变化的影响及各部门产量变化对最终产品变化的影响.

3. 对计划中重大项目建设的平衡计算

(1)项目建设时的平衡计算.将计划建设项目(包括附属工程)所需消耗的产品视为对最终产品需求的增加,采用公式 $X^* = (I-A)^{-1}Y^*$ 计算项目建设时所需各部门的产量,其中 $Y^* = (Y_1, \cdots, Y_K + \Delta Y_K, \cdots, Y_L + \Delta Y_L, \cdots, Y_n)^{\mathrm{T}}$,$\Delta Y_K, \cdots, \Delta Y_L$ 为计划建设项目对 $K, \cdots, L$ 部门产品的需求量.

(2)项目建成投产后的平衡计算.假设某一项目投产后仅使第 n 部门产量增加 ΔX_n,这样会对其他 $n-1$ 个部门产生直接消耗和间接消耗的需求,使这 $n-1$ 个部门的产量需求发生变化,其公式为

$$(\Delta X_1, \Delta X_2, \cdots, \Delta X_{n-1})^{\mathrm{T}} = (I - A_{n-1})^{-1}(a_{1n}, a_{2n}, \cdots, a_{n-1,n})^{\mathrm{T}}\Delta X_n, \quad (1\text{-}61)$$

其中 $A_{n-1} = \begin{bmatrix} a_{11} & \cdots & a_{1,n-1} \\ \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n-1} \end{bmatrix}$,$I$ 为 $n-1$ 阶单位阵.

4. 与数学规划结合,制订出国民经济发展的最优方案

可与线性规划、非线性规划、动态规划等最优化方法结合起来,在一定资源、技术的条件下制订合理的计划,以取得最佳的经济效果,或在一定的经济效果前提下寻求消耗最少的计划方案等.具体方法在本篇第6章介绍.

2 静态产品投入产出表的编制方法

投入产出模型是建立在投入产出表的基础上的,投入产出表的编制需要解决一系列的分类、组合、修订和平衡等问题.

2.1 静态产品投入产出表编制的基本问题

1. 部门划分问题

该问题是编制投入产出表首先遇到的一个问题.从国民经济平衡和经济分析、预测、计划与政策研究等不同需要看,不可能也不必要把成千上万种社会产品都进行具体的分析,只需将它们组合成若干个大类和部门进行分析和研究.部门划分合适与否直接影响着投入产出法的应用效果,划分时须以社会劳动分工的发展水平为依据,同时须遵循一定的分类原则,且要根据编表目的,结合国家、地区等经济特点来进行.

(1)部门划分原则.部门,是根据产品同类性原则组成的同类性产品的综合体.所谓**同类性产品**是指消耗结构相同、工艺技术相同和经济用途相同的产品.上述原

则是为了使消耗系数能够较准确地反映部门间的技术经济联系，保证投入和产出间的线性关系.实际上，部门划分的同类性原则中三个条件并不协调一致，部门划分越详细，三个条件越趋于协调；反之，三个条件则相差越大.部门个数的多少须同时依据编表的目的、收集数据的可能来确定.一般而言，若侧重于理论分析和政策研究，可以粗一些；若主要用于短期计划和生产组织，则可以细一些，如实物型表中部门可在 100~200 个左右，价值型表中部门可在 50~150 个左右，等等.

(2)物质生产部门和非物质生产部门的划分.马克思主义经济学将国民经济划分为物质生产部门和非物质生产部门.凡直接创造物质产品或增加产品价值的部门为物质生产部门，如农业、工业、建筑业、商业、饮食业及服务于生产的交通运输业，而教育、客运业、卫生和服务于居民的文艺、邮政业等属于非物质生产部门.

1)物质生产部门的产品部门划分.根据产品同类性原则将物质生产部门进行划分所得的部门称为**产品部门**.它不同于企业部门(经济部门)和行政部门.依据企业主要产品，以企业为单位按照企业的同类性划分的部门为企业部门或经济部门.企业除主要产品外，可能还生产许多次要产品，这些次要产品分属不同的产品部门.产品部门与按行政管理系统划分的行政主管部门(如部、厅、局等)差别更大.

2)非物质生产部门的劳务部门划分

非物质生产部门给社会提供的劳务产品同样具有价值和使用价值.**劳务部门**可分为两大类：一类是商品性劳务部门，是盈利的非物质生产部门，如电影院、客运、旅店等；另一类是非商品性劳务部门，是非盈利的非物质生产部门，如科研、卫生部门、机关团体等.

2.产值计算问题

产值有不同的计算方法，各方法的计算结果会有较大差异.

(1)产值计算方法

1)部门法.部门法以部门为单位，部门的总产值不包含部门内部各种生产间的相互消耗，即部门总产值不受部门内部组织结构或部门内各企业间分合变动的影响.因此，采用部门法计算总产值时，投入产出表第Ⅰ象限的主对角线全为 0.

2)工厂法.工厂法以工厂企业为单位，企业总产值不包含企业内部各种生产间的相互消耗，只计算企业对外销售的产值.用工厂法计算的部门总产值一般比部门法计算的部门总产值大.

3)产品法.产品法以产品为单位，在生产过程中所有周转额都计入产值，即同一部门或企业内部生产间的相互消耗都计入产值，因此，按产品法计算的部门总产值一般要比工厂法计算的部门总产值大.

上述三种产值计算方法各有利弊，从减少产值重复计算看，部门法最好，工厂法次之，产品法最差；从全面反映所有产品周转活动及其相互联系看，产品法最好，部门法最差.

在现行统计中，鉴于农业生产过程相对简单，内部联系也较少，故农业产值和产量按产品法计算.因工业生产过程较复杂，相互联系紧密，若工业产值按工厂法计算，产量按产品法计算，则两种计算方法会造成工业产值和产量上的差异，因此我国一般采用现行统计制度的计算方法，即农业用产品法，工业用工厂法.当然，若

条件允许,可以同时给出产品法和工厂法两套数据,以利于比较和分析.

(2)非物质生产部门劳务总值的计算.非物质生产部门劳务总值,按劳务性质,分两种情况进行计算:一是商品性劳务,按其价值总额(即劳务的销售收入)计算总产值,如保险业按保险费、手续费与赔偿金之差计算产值,银行业按利息收入净额、业务收入和手续费计算产值等;二是非商品性劳务,按其劳务成本(即费用总额)计算总产值,这类劳务是非盈利的,总产值中不包含利润,大部分情况下也不包含间接税.

劳务总值计算公式为

劳务总值 = 产品和劳务的中间投入 + 固定资产折旧 + 工资 +
福利基金 + (利润) + (间接税) + 利息 + 其他.

3. 价格问题

编制价值型投入产出表必须采用统一的价格体系.在现实经济活动中存在着多种不同的价格体系.

(1)计划价格与协议价格.**计划价格**为由国家各级政府有计划规定的价格,又称平价,可分为中央管理价格、地方管理价格及有限制的浮动价格.**协议价格**为由买卖双方经协商议定的价格,常称议价,包括企业自销价格、集市贸易价格及外贸价格等.

(2)生产者价格与消费者价格.**生产者价格**是在直接生产过程中形成的价格,包括工业品的出厂价和农产品的采购价.**消费者价格**是消费者在市场上采购时的价格,即批发零售价格,它包含直接生产费用及作为生产过程在流通领域内的继续而引起的运输费用和商业费用.采用两种价格编表,各有利弊,采用生产者价格能较准确地反映部门间的生产技术联系,但资料不易取得,不能全面反映现实的国民经济周转;采用消费者价格能较全面地反映国民经济的周转,资料容易取得,但不能确切反映部门间的生产技术联系.

(3)不变价格与现行价格.**不变价格**指考虑利息、物价上涨等因素将现行价格转化为某一参照时间的价格,使不同时期的价格具有可比性,但不变价格不利于如实反映以货币表示的生产水平.现行价格能够如实反映以货币表示的生产水平,但它不能反映实际产量等变化情况,不同时期的现行价格不具有可比性.

(4)国内价格与国际价格.国际价格,就进口而言,为到岸价格,或再加上进口税,不包括国内运输费用和流通费用;就出口而言,为离岸价格.

从国内编表看,多数采用国内现行的生产者综合价格,即国内价格、现行价格、生产者价格和计划价格.但在资料难以取得的情况下,亦可考虑采用其他价格.

4. 固定资产折旧与更新、进出口等问题

(1)固定资产折旧与更新.固定资产在生产中消耗的价值是以基本折旧和大修理折旧的形式经逐步提取的方式而补偿的,固定资产的实物更新是以大修理和更新的形式实现的,这就使当年提取的折旧额与更新所需数额往往不一致.此外,更新常常伴随着改造,不会仅在原有技术水平上重置.因此,正确处理固定资产折旧与更新,是编制投入产出表时必须认真对待的一个问题.

在实物型投入产出表中,纵列因计量单位不同而不能相加,所以固定资产的消

耗不能以折旧的形式在纵列栏中体现，而只能在横行的最终产品栏内单列一项“固定资产更新、改造和大修理”，以便与部门产品（主要是机械产品、建筑安装工程产品）用于积累的部分相区别，通常，这方面数据较难取得，同时更新（简单再生产）与积累也不易区别.

在价值型投入产出表中，一般的处理方式是，将固定资产折旧列入第Ⅲ象限，将更新、改造和大修理列入第Ⅱ象限，但对固定资产更新、改造和大修理的处理有以下几种方式.

1)从实际补偿考虑.物质生产部门提取的基本折旧和大修理折旧，形成固定资产简单再生产专用基金，其中一部分真正用于固定资产的更新、改造和大修理，一部分用于基本建设，还有一部分用于增加流动资产.只将实际用于固定资产补偿的部分列入“固定资产更新、改造和大修理”中，而将其余部分列入当年积累中.

2)从应有用途考虑.不论固定资产简单再生产专用基金用于何处，一律处理为固定资产的更新、改造和大修理，并与折旧提取额相等.这种方法固然简单，但不能反映实际补偿情况.

3)与固定资产积累合并.将固定资产更新、改造和大修理与当年积累合并在一起计算，形成固定资产的总投资.同样，这种方法也无法反映实际补偿情况，从而无法区分固定资产补偿和新增固定资产情况.

(2)进口与出口（调入与调出）.进出口情况可采用以下方法处理.

1)差额法.对进出口商品按主栏部门分类，将其差额（出口 - 进口）作为一列置于第Ⅱ象限.该方法无法反映进口产品的分配和使用情况，该方法只简单地认为各部门进口产品的分配系数与国内的一致.显然，上述假定与实际不符，当进口产品与国内产品的价格不一致时，差距就更大.此外，进口产品的间接消耗发生在国外，上述处理方式计算所得的完全消耗系数也无法准确反映国内产品的完全消耗，因此，差额法只适用于进出口比重很小的情况.

2)向量法.将进口商品按宾栏部门分类，单独列出进口商品的分配使用情况，作为一个行向量置于Ⅰ、Ⅱ象限与Ⅲ、Ⅳ象限之间，而出口仍作为一个列向量置于第Ⅱ象限.这种处理是假定各部门进口产品的直接消耗系数与国内的一致，即假定同一部门消耗的各种产品中，进口产品所占的比重与国内的一致.这种假设也有不足.此外，从Ⅰ、Ⅱ象限中排除进口产品的同时，也排除了进口替代（以国内商品代替进口商品）的可能性.

3)矩阵法.对所有进口商品按主、宾栏分类，形成一个流量矩阵.采用矩阵法能够对进口商品进行清楚而准确的分析，但需掌握进口商品的详细资料，增加了搜集数据与计算的工作量和难度.

地区投入产出表中产品的调入、调出与全国投入产出表的进出口有类似的性质，详细情况将在本篇第3章介绍.

2.2 直接消耗系数的修订和预测

直接消耗系数反映的是部门间的技术经济联系，一般假定在较短的时期内大

多数系数相对稳定.实际上,这些系数不是固定不变的,尤其当计划期较长时.为使其适应于计划期生产技术变化的情况,可采取下述一些修订方法对其进行修订.

1.调查研究法

此法邀请各有关方面的专家,针对生产技术变化的实际情况和预测结果,共同研究与确定计划期的直接消耗系数.此方法简单,耗费资金和时间均较少,其结果有一定可信度.

2.定期编制修订法

根据基层和统计的实际资料,通过定期编制投入产出表进行修订.这种方法一般需投入一定的资金和时间去搜集、整理、分析数据资料,其结果具有较高的可信度.

3.适时修正法

此方法亦称 RAS 法,适于对价值型投入产出表直接消耗系数进行修订.主要根据计划期的总产量、最终产品量及新创造价值的数额等得到计划期各部门中间产品的合计值和物质消耗的合计值,以此来调整报告期的直接消耗系数,得到计划期的直接消耗系数.具体方法如下.

假设报告期各部门中间产品合计值组成的列向量和物质消耗合计值组成的行向量分别为 $\boldsymbol{u}_0=(u_{01},u_{02},\cdots,u_{0n})^{\mathrm{T}}$ 和 $\boldsymbol{v}_0=(v_{01},v_{02},\cdots,v_{0n})$,计划期各部门中间产品的合计值组成的列向量和物质消耗合计值组成的行向量分别为 $\boldsymbol{u}_0^*=(u_{01}^*,u_{02}^*,\cdots,u_{0n}^*)^{\mathrm{T}}$ 和 $\boldsymbol{v}_0^*=(v_{01}^*,v_{02}^*,\cdots,v_{0n}^*)$,报告期直接消耗系数为 $\boldsymbol{A}$,第Ⅰ象限流量矩阵为 $\boldsymbol{X}=(\overline{x}_{ij})_{n\times n}$,记调整后第Ⅰ象限的流量矩阵为 $\boldsymbol{X}^*=(\overline{x}_{ij}^*)_{n\times n}$,则 $\sum_{j=1}^{n}\overline{x}_{ij}=u_{0i}$, $\sum_{i=1}^{n}\overline{x}_{ij}=v_{0j},i,j=1,2,\cdots,n.$

调整过程为:

第1步 令 $\boldsymbol{X}^{(1)}=(x_{ij}^{(1)})_{n\times n}=\boldsymbol{RX},u_{0j}^{(1)}=u_{0j}^*,v_{0j}^{(1)}=\sum_{i=1}^{n}x_{ij}^{(1)},j=1,2,\cdots,n,$

其中 $\boldsymbol{R}=\mathrm{diag}(r_1,r_2,\cdots,r_n),r_i=u_{0i}^*/u_{0i},i=1,2,\cdots,n.$

第2步 令 $\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}\boldsymbol{S},v_{0j}^{(2)}=v_{0j}^*,u_{0j}^{(2)}=\sum_{i=1}^{n}x_{ji}^{(2)},j=1,2,\cdots,n,$

其中 $\boldsymbol{S}=\mathrm{diag}(s_1,s_2,\cdots,s_n),s_j=v_{0j}^*/v_{0j}^{(1)},j=1,2,\cdots,n.$

第3步 计算 $W=\sum_{i=1}^{n}(u_{0i}^{(2)}-u_{0i}^{(1)})^2+\sum_{j=1}^{n}(v_{0j}^{(2)}-v_{0j}^{(1)})^2,$

若 $W<\varepsilon$(ε 为事先给定的相当小的正数),则令 $\boldsymbol{X}^*=\boldsymbol{X}^{(2)}$,由 $\boldsymbol{X}^*$ 和计划期各部门的总产值计算出计划期的直接消耗系数,计算结束;否则,令 $u_{0i}=u_{0i}^{(2)},i=1,2,\cdots,n,\boldsymbol{X}=\boldsymbol{X}^{(2)}$,返回第1步.

4.修正的 RAS 法

在 RAS 法的运算过程中,只要流量矩阵 $\boldsymbol{X}$ 的某一元素为0,无论怎么调整,该

元素始终为0.利用上述特点,可把RAS法与其他方法结合起来修订直接消耗系数.例如,可经调查核实,确定 X 中的一批数据,在运用RAS法时,先令这些已确定的数据为0,重新计算得到相应的 u_0, v_0, u_0^* 和 v_0^*,利用RAS法对经上述处理过的数据进行调整,调整结束后,再将原已核实确定的数据还原,进一步可得到计划期的直接消耗系数.

2.3 编制投入产出表的推导法

编制投入产出表的推导法亦称 **UV 表法**,该方法直接将基层企业的原始数据编成"产品"和"部门"混合型投入产出表.根据一定的假定,采用数学方法推导出"纯"部门的投入产出系数,再编制投入产出表.这里所说的"产品"是指产品部门或纯部门,"部门"是指企业部门.

1."产品"与"部门"投入产出表及其模型

(1)"产品"与"部门"投入产出表的结构.表2-1给出了"产品"与"部门"投入产出表的结构.为保证数学建模运算的可行性和正确性,要求在表2-1中,产品和部门的分类按一定方式进行,亦即就企业主要产品(特征产品)而言,两者的分类安排应基本对应一致,也就是说要求第 i 产品部门的产品与第 i 企业部门的主要产品是一致的.表2-1中,X 阵(消耗阵、投入阵或U表),其元素 x_{ij} 表示 j 部门中 i 产品的投入量,或 j 部门消耗 i 产品的数量,S 阵(制造阵、产出阵或V表),其元素 s_{ij} 表示 i 部门生产 j 产品的数量,或 j 产品由 i 部门生产的数量.

表2-1

		产品				部门				最终产品	总产出
		1	2	…	n	1	2	…	m		
产品	1					x_{11}	x_{12}	…	x_{1m}	Y_1	Z_1
	2					x_{21}	x_{22}	…	x_{2m}	Y_2	Z_2
	⋮					⋮	⋮		⋮	⋮	⋮
	n					x_{n1}	x_{n2}	…	x_{nm}	Y_n	Z_n
部门	1	s_{11}	s_{12}	…	s_{1n}						G_1
	2	s_{21}	s_{22}	…	s_{2n}						G_2
	⋮	⋮	⋮		⋮						⋮
	m	s_{m1}	s_{m2}	…	s_{mn}						G_m
新创造价值						F_1	F_2	…	F_m		W
总产值		Z_1	Z_2	…	Z_n	G_1	G_2	…	G_m		

(2)"产品"与"部门"投入产出模型.

$$\sum_{j=1}^{m} x_{ij} + Y_i = Z_i, \quad i = 1,2,\cdots,n, \tag{2-1}$$

$$\sum_{i=1}^{m} s_{ij} = Z_j, \quad j = 1,2,\cdots,n, \tag{2-2}$$

$$\sum_{j=1}^{n} s_{ij} = G_i, \quad i = 1,2,\cdots,m, \tag{2-3}$$

$$\sum_{i=1}^{n} x_{ij} + F_j = G_j, \quad j = 1,2,\cdots,m. \tag{2-4}$$

因此,

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} + \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} Z_i, \tag{2-5}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij} = \sum_{i=1}^{m} G_i, \tag{2-6}$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij} + \sum_{j=1}^{m} F_j = \sum_{j=1}^{m} G_j, \tag{2-7}$$

$$\sum_{i=1}^{n} Y_i = \sum_{j=1}^{m} F_j. \tag{2-8}$$

(3)由产品与部门投入产出表导出的结构系数.产品与部门投入产出表可以提供如下三种结构系数,即部门对产品的消耗系数、部门生产各种产品的比例系数及产品的部门份额.

消耗系数阵 $B = (b_{ij})_{n\times m} = X\widehat{G}^{-1}$, (2-9)

比例系数阵 $C = (c_{ij})_{n\times m} = S^{\mathrm{T}}\widehat{G}^{-1}$, (2-10)

部门份额矩阵 $D = (d_{ij})_{m\times n} = S\widehat{Z}^{-1}$, (2-11)

其中

$$\widehat{G} = \mathrm{diag}(G_1, G_2, \cdots, G_m), \widehat{Z} = \mathrm{diag}(Z_1, Z_2, \cdots, Z_n);$$

b_{ij}为j部门生产单位产品消耗i种产品的数量;c_{ij}为j部门生产的产品i在该部门产品总量中的比重;d_{ij}为产品j的总量中由i部门生产的部分所占的比重.

2.产品与产品投入产出表及其模型

(1)基本假定.一是产品工艺假定,即同一种产品,无论由哪个部门生产,都有相同的消耗构成;二是部门工艺假定,即同一企业部门,无论生产哪种产品,其消耗构成相同.这两种工艺假定一般不能同时成立.

(2)产品与产品投入产出表的编制.

1)在产品工艺假定下产品间的消耗构成.在产品工艺假定下,同一种产品无论由哪个部门生产都有相同的消耗构成,因此,j部门单位产值对i产品的直接消耗系数b_{ij},是j部门生产各种产品的单位产值对i产品直接消耗系数$a_{i1}^c, a_{i2}^c, \cdots, a_{in}^c$的加权平均值,其权数是$j$部门生产的各种产品在$j$部门总产值中所占的比重$c_{1j}$, $c_{2j}, \cdots, c_{nj}$,即

或

$$b_{ij} = \sum_{k=1}^{n} a_{ik}^{c} c_{kj} \quad (i = 1,2,\cdots,n; j = 1,\cdots,m), \tag{2-12}$$

$$\boldsymbol{B} = \boldsymbol{A}_c \boldsymbol{C}, \tag{2-13}$$

其中 $\boldsymbol{A}_c = (a_{ij}^{c})_{n\times n}$为在产品工艺假定下的产品与产品的直接消耗系数阵.

2)在部门工艺假定下产品间的消耗构成.在部门工艺假定下,同一部门生产的各种产品都有相同的消耗构成,因此,j 部门单位产值对 i 产品的消耗 b_{ij}即为该部门生产单位产品对i 产品的消耗.由于各部门对 i 产品的消耗系数不同,因此,单位 j 产品对 i 产品的消耗量 a_{ij}^{D}应是各部门生产单位产品对 i 产品消耗量 b_{ik}, $k = 1,2,\cdots,n$ 的加权平均数,其权数是 j 产品总量中各部门所占的比重,即

$$a_{ij}^{D} = \sum_{k=1}^{n} b_{ik} d_{kj}, \quad i,j = 1,2,\cdots,n, \tag{2-14}$$

或

$$\boldsymbol{A}_D = \boldsymbol{B}\boldsymbol{D}, \tag{2-15}$$

其中 $\boldsymbol{A}_D = (a_{ij}^{D})_{n\times n}$为在部门工艺假定下产品与产品的直接消耗系数阵.

3)产品与产品投入产出表的编制.

第 1 步　计算按部门划分的原始投入

$$\boldsymbol{F}_D = \boldsymbol{F}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{D}\widehat{\boldsymbol{Z}} \quad 或 \quad \boldsymbol{F}_c = \boldsymbol{F}\widehat{\boldsymbol{G}}^{-1}\boldsymbol{C}^{-1}\widehat{\boldsymbol{Z}}, \tag{2-16}$$

当 $n \neq m$ 时,$\boldsymbol{C}^{-1}$不存在,$\boldsymbol{F}_c$ 不可用.

第 2 步　由 $\boldsymbol{A}_c$ 或$\boldsymbol{A}_D$ 建立Ⅰ,Ⅲ象限平衡方程式

$$\boldsymbol{A}_c^{*}\boldsymbol{Z} + \boldsymbol{F}_c = \boldsymbol{Z} \quad 或 \quad \boldsymbol{A}_D^{*}\boldsymbol{Z} + \boldsymbol{F}_D = \boldsymbol{Z}, \tag{2-17}$$

其中 $\boldsymbol{A}_c^{*} = \mathrm{diag}\left(\sum_{i=1}^{n} a_{i1}^{c}, \sum_{i=1}^{n} a_{i2}^{c}, \cdots, \sum_{i=1}^{n} a_{in}^{c}\right)$, $\boldsymbol{A}_D^{*} = \mathrm{diag}\left(\sum_{i=1}^{n} a_{i1}^{D}, \sum_{i=1}^{n} a_{i2}^{D}, \cdots, \sum_{i=1}^{n} a_{in}^{D}\right)$.

第 3 步　建立Ⅰ,Ⅱ象限平衡方程式

$$\boldsymbol{A}_c\boldsymbol{Z} + \boldsymbol{Y} = \boldsymbol{Z} \quad 或 \quad \boldsymbol{A}_D\boldsymbol{Z} + \boldsymbol{Y} = \boldsymbol{Z}. \tag{2-18}$$

3. 部门间的消耗构成

(1)基本假定.一是部门份额假定,即每一产品由各部门生产的比例份额保持不变;二是产品比例假定,即同一部门生产的各种产品间的比例保持不变.

(2)在部门份额假定下部门间的消耗构成.在部门份额假定下,每一产品的产出在各部门的份额保持不变,j 部门生产单位产品对 i 部门的消耗 e_{ij}^{D},是 j 部门生产单位产品对各种产品消耗系数 b_{kj}($k = 1,2,\cdots,n$)的加权平均数,其权重为 i 部门各种产品的份额 D_{ik}, $k = 1,2,\cdots,n$,即

$$e_{ij}^{D} = \sum_{k=1}^{n} D_{ik} b_{kj}, \quad i,j = 1,2,\cdots,m, \tag{2-19}$$

或

$$\boldsymbol{E}_D = \boldsymbol{D}\boldsymbol{B}, \tag{2-20}$$

其中

$$E_D = (e_{ij}^D)_{m \times m}.$$

(3)在产品比例假定下部门间的消耗构成.在产品比例假定下,同一部门生产的各种产品,其比例保持不变.j部门单位产品消耗i产品的系数b_{ij}是j部门单位产品消耗各部门产品的系数$e_{kj}^c(k=1,2,\cdots,m)$的加权平均值,其权重为i产品在各部门产品中的比例,即

$$b_{ij} = \sum_{k=1}^{m} c_{ik} e_{kj}^c, \quad i = 1,2,\cdots,n; \quad j = 1,2,\cdots,m, \tag{2-21}$$

或记为

$$B = CE_c, \tag{2-22}$$

其中 $E_c = (e_{ij}^c)_{m \times m}$.

3 地区、地区间投入产出模型

3.1 地区投入产出模型

地区投入产出模型,一般是指根据全国在地域上的一个组成部分编制的投入产出模型,主要指根据行政区域编制的模型.地区投入产出模型的研制,有助于了解地区经济全貌和产业特点,有助于了解本地区各部门间和本地区与其他地区间的技术经济联系,有助于提高地区计划工作水平,加强地区综合平衡,同时还起到丰富全国投入产出表的内容、了解某项经济政策对地区经济带来的影响等作用.

1.地区投入产出模型的特点

(1)调入、调出量占地区总产量的比重较大.由于地区经济远非一个完整的体系,地区之间存在着较多的互补性联系,因此,同全国性投入产出表相比,调入、调出量在地区经济中占较大比重.

(2)部门划分有粗细之分.一般而言,不同地区经济各有侧重、各具特点,为较好地反映地区经济这一特征,需要将地区经济中某些重点和特色部门划分得细些,将次要的部门划分得粗些.

2.地区投入产出模型

(1)地区投入产出表的模式.根据对调入、调出的不同处理,在抽象化处理固定资产折旧的条件下,可将地区静态产品投入产出表分为三种模式.

1)简单模式.如表3-1和表3-2所示.表3-1表示地区实物型投入产出表,表3-2表示地区价值型投入产出表.

在表3-1、表3-2中,只把调入、调出作为列向量列入最终产品栏中,类似于全国投入产出表对进出口处理的差额法.

2)一般模式.此模式如表3-3和表3-4所示.表3-3为地区实物型投入产出表的一般模式,表3-4为地区价值型投入产出表的一般模式.

表 3-1

<table>
<tr><th rowspan="4">投入</th><th colspan="16">产　出</th></tr>
<tr><th rowspan="3">计量单位</th><th colspan="5">中间产品</th><th colspan="8">最终产品</th><th rowspan="3">合计 Y_i</th><th rowspan="3">总产出 Q_i</th></tr>
<tr><th rowspan="2">1</th><th rowspan="2">2</th><th rowspan="2">…</th><th rowspan="2">n</th><th rowspan="2">合计</th><th colspan="3">消　费</th><th colspan="3">积　累</th><th rowspan="2">调出(+)</th><th rowspan="2">调入(-)</th></tr>
<tr><th>个人消费</th><th>社会消费</th><th>小计</th><th>新增固定资产</th><th>增加库存</th><th>小计</th></tr>
<tr><td>1</td><td></td><td>q_{11}</td><td></td><td>…</td><td>q_{1n}</td><td rowspan="4"></td><td colspan="6" rowspan="4">Ⅱ</td><td colspan="2" rowspan="4">Ⅲ</td><td>Y_1</td><td>Q_1</td></tr>
<tr><td>2</td><td></td><td colspan="4">Ⅰ</td><td>Y_2</td><td>Q_2</td></tr>
<tr><td>⋮</td><td></td><td>⋮</td><td></td><td></td><td>⋮</td><td>⋮</td><td>⋮</td></tr>
<tr><td>n</td><td></td><td>q_{1n}</td><td></td><td>…</td><td>q_{nn}</td><td>Y_n</td><td>Q_n</td></tr>
</table>

表 3-2　　单位:货币单位

<table>
<tr><th colspan="2" rowspan="4">投　入</th><th colspan="15">产　出</th></tr>
<tr><th colspan="5">中间产品</th><th colspan="9">最终产品(Y)</th><th rowspan="3">总计 X</th></tr>
<tr><th rowspan="2">1</th><th rowspan="2">2</th><th rowspan="2">…</th><th rowspan="2">n</th><th rowspan="2">合计</th><th colspan="3">地区内 (Y_D)</th><th rowspan="2">调出 (F)</th><th rowspan="2">调入 ($-G$)</th><th rowspan="2">出口 (O)</th><th rowspan="2">进口 ($-P$)</th><th rowspan="2">合计</th></tr>
<tr><th>消费</th><th>积累</th><th>合计</th></tr>
<tr><td>物资消耗 C</td><td>部门 1
部门 2
⋮
部门 n
合　计</td><td colspan="5">Ⅰ</td><td colspan="9">Ⅱ</td><td></td></tr>
<tr><td>新创造价值 N</td><td>劳动报酬(V)
社会纯收入(M)
合　计</td><td colspan="5">Ⅲ</td><td colspan="9">Ⅳ</td><td></td></tr>
<tr><td colspan="2">总产值(X)</td><td colspan="15"></td></tr>
</table>

表 3-3

<table>
<tr><td colspan="2" rowspan="4">投　　入</td><td colspan="20">产　　出</td></tr>
<tr><td rowspan="3">计量单位</td><td colspan="5">中间产品</td><td colspan="13">最终产品</td><td rowspan="3">总产出</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">2</td><td rowspan="2">…</td><td rowspan="2">n</td><td rowspan="2">合计</td><td rowspan="2">固定资产更改大修</td><td colspan="3">消　费</td><td colspan="3">积　累</td><td rowspan="2">调入(－)</td><td rowspan="2">调出(＋)</td><td rowspan="2">进口(－)</td><td rowspan="2">出口(＋)</td><td rowspan="2">合计</td></tr>
<tr><td>个人消费</td><td>社会消费</td><td>小计</td><td>新增固定资产</td><td>增加库存</td><td>小计</td></tr>
<tr><td rowspan="4">本地生产</td><td>1</td><td colspan="6">q^d_{11} ··· q^d_{1n}</td><td colspan="7">y^d_{11} ··· y^d_{1m}</td><td>O_1</td><td>E_1</td><td>F_1</td><td>G_1</td><td>Y_1</td><td>X_1</td></tr>
<tr><td>2</td><td colspan="6"></td><td colspan="7"></td><td>O_2</td><td>E_2</td><td>F_2</td><td>G_2</td><td>Y_2</td><td>X_2</td></tr>
<tr><td>⋮</td><td colspan="6">⋮ Ⅰ ⋮</td><td colspan="7">⋮ Ⅱ ⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>Ⅲ ⋮</td><td>⋮</td><td>⋮</td></tr>
<tr><td>n</td><td colspan="6">q^d_{n1} ··· q^d_{nn}</td><td colspan="7">y^d_{n1} ··· y^d_{nm}</td><td>O_n</td><td>E_n</td><td>F_n</td><td>G_n</td><td>Y_n</td><td>X_n</td></tr>
<tr><td rowspan="4">外地调入</td><td>1</td><td colspan="6">q^h_{11} ··· q^h_{1n}</td><td colspan="7">y^h_{11} ··· y^h_{1m}</td><td>H_1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td colspan="6"></td><td colspan="7"></td><td>H_2</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>⋮</td><td colspan="6">⋮ Ⅳ ⋮</td><td colspan="7">⋮ Ⅴ ⋮</td><td>⋮</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>n</td><td colspan="6">q^h_{n1} ··· q^h_{nn}</td><td colspan="7">y^h_{n1} ··· y^h_{nm}</td><td>H_n</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

表 3-4　　　　单位：货币单位

<table>
<tr><td colspan="2" rowspan="4">投　　入</td><td colspan="20">产　　出</td></tr>
<tr><td rowspan="3">计量单位</td><td colspan="5">中间产品</td><td colspan="13">最终产品</td><td rowspan="3">总产出</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">2</td><td rowspan="2">…</td><td rowspan="2">n</td><td rowspan="2">合计</td><td rowspan="2">固定资产更改大修</td><td colspan="3">消　费</td><td colspan="3">积　累</td><td rowspan="2">调入(－)</td><td rowspan="2">调出(＋)</td><td rowspan="2">进口(－)</td><td rowspan="2">出口(＋)</td><td rowspan="2">合计</td></tr>
<tr><td>个人消费</td><td>社会消费</td><td>小计</td><td>新增固定资产</td><td>增加库存</td><td>小计</td></tr>
<tr><td rowspan="5">本地生产</td><td>1</td><td colspan="6">x^d_{11} ··· x^d_{1n}</td><td colspan="7">y^d_{11} ··· y^d_{1m} y_{D_1}</td><td>O_1</td><td>E_1</td><td>F_1</td><td>G_1</td><td>Y_1</td><td>X_1</td></tr>
<tr><td>2</td><td colspan="6"></td><td colspan="7"></td><td>O_2</td><td>E_2</td><td>F_2</td><td>G_2</td><td>Y_2</td><td>X_2</td></tr>
<tr><td>⋮</td><td colspan="6">⋮ Ⅰ ⋮</td><td colspan="7">⋮ Ⅱ ⋮ ⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>Ⅲ ⋮</td><td>⋮</td><td>⋮</td></tr>
<tr><td>n</td><td colspan="6">x^d_{n1} ··· x^d_{nn}</td><td colspan="7">y^d_{n1} ··· y^d_{nm} y_{D_n}</td><td>O_n</td><td>E_n</td><td>F_n</td><td>G_n</td><td>Y_n</td><td>X_n</td></tr>
<tr><td>合计</td><td colspan="6"></td><td colspan="7"></td><td colspan="5"></td><td></td></tr>
<tr><td rowspan="5">外地调入</td><td>1</td><td colspan="6">x^h_{11} ··· x^h_{1n}</td><td colspan="7">y^h_{11} ··· y^h_{1m}</td><td>H_1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td colspan="6"></td><td colspan="7"></td><td>H_2</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>⋮</td><td colspan="6">⋮ Ⅳ ⋮</td><td colspan="7">⋮ Ⅴ ⋮</td><td>⋮</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>n</td><td colspan="6">x^h_{n1} ··· x^h_{nn}</td><td colspan="7">y^h_{n1} ··· y^h_{nm}</td><td>H_n</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>合计</td><td colspan="6"></td><td colspan="7"></td><td colspan="5"></td><td></td></tr>
</table>

投入		产出																		
		计量单位	中间产品					最终产品												总产出
			1	2	…	n	合计	固定资产更改大修	消费			积累			调入(−)	调出(+)	进口(−)	出口(+)	合计	
									个人消费	社会消费	小计	新增固定资产	增加库存	小计						
固定资产折旧			$D_1 \cdots D_n$																	
新创造价值	劳动报酬		$V_1 \cdots V_n$					Ⅶ												
	社会纯收入		Ⅵ $m_1 \cdots m_n$																	
	合计																			
总产值			$X_1 \cdots X_n$																	

由表 3-3 和表 3-4 可见，地区投入产出表的一般模式相当于全国投入产出表对进出口处理的矩阵法．

3)综合模式．此模式如表 3-5 和表 3-6 所示．表 3-5 为地区实物型投入产出表的综合模式，表 3-6 为地区价值型投入产出表的综合模式．

表 3-5

输入地区			投入		产出																
					计量单位	中间产品					最终产品										总产出
											本地使用				输出地区						
1 2 … k	进口	合计				1	2	…	n	合计	固定资产更新改造和大修理	消费	积累	合计	1	2	…	k	出口	合计	
			本地生产	1 2 … n		Ⅰ					Ⅱ				Ⅲ						
Ⅳ			外地输入	1 2 … n		Ⅴ					Ⅵ										

表 3-6 单位:货币单位

<table>
<tr><td colspan="6" rowspan="4"></td><td colspan="2" rowspan="4">投 入</td><td colspan="16">产 出</td></tr>
<tr><td colspan="5">中间产品</td><td colspan="10">最终产品</td><td rowspan="3">总产出</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">2</td><td rowspan="2">…</td><td rowspan="2">n</td><td rowspan="2">合计</td><td colspan="4">本地使用</td><td colspan="6">输出地区</td></tr>
<tr><td>固定资产更新改造和大修理</td><td>消费</td><td>积累</td><td>合计</td><td>1</td><td>2</td><td>…</td><td>k</td><td>出口</td><td>合计</td></tr>
<tr><td colspan="6">输入地区</td><td rowspan="3">本地生产</td><td rowspan="2">1
2
⋮
n</td><td colspan="5" rowspan="3">Ⅰ</td><td colspan="3" rowspan="3">Ⅱ</td><td rowspan="3"></td><td colspan="4" rowspan="3">Ⅲ</td><td rowspan="3"></td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td>1</td><td>2</td><td>…</td><td>k</td><td>进口</td><td>合计</td></tr>
<tr><td colspan="6"></td><td>合计</td></tr>
<tr><td colspan="4" rowspan="2">Ⅳ</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2">外地输入</td><td>1
2
⋮
n</td><td colspan="5" rowspan="2">Ⅴ</td><td colspan="3" rowspan="2">Ⅵ</td><td rowspan="2"></td><td colspan="7" rowspan="2"></td></tr>
<tr><td>合计</td></tr>
<tr><td colspan="6" rowspan="4"></td><td colspan="2">固定资产折旧</td><td colspan="5"></td><td colspan="3"></td><td></td><td colspan="7" rowspan="5"></td></tr>
<tr><td rowspan="3">新创造价值</td><td>劳动报酬</td><td colspan="5" rowspan="3">Ⅶ</td><td colspan="4" rowspan="3">Ⅷ</td></tr>
<tr><td>纯收入</td></tr>
<tr><td>合计</td></tr>
<tr><td colspan="6"></td><td colspan="2">总产值</td><td colspan="5"></td><td colspan="4"></td></tr>
</table>

由表 3-5 和表 3-6 可见,综合模式将输入、输出地区详细分列,可全面反映地区调入、调出情况.

(2) 地区投入产出模型

以表 3-2 为例,其模型为

$$AX + Y_D + F - G + O - P = X, \tag{3-1}$$

或

$$X = (I - A)^{-1}(Y_D + F - G + O - P). \tag{3-2}$$

$$A_c X + V + M = X, \tag{3-3}$$

其中 $A = (a_{ij})_{n\times n}$ 为直接消耗系数矩阵, $A_c = \mathrm{diag}(\sum_{i=1}^{n} a_{i1}, \sum_{i=1}^{n} a_{i2}, \cdots, \sum_{i=1}^{n} a_{in})$.

现考察调入产品、进口产品对区内各部门产品产量的影响,并以此了解调入产品、进口产品与本地区生产之间的联系,为此,定义调入系数 t_i 和进口系数 s_i 分别为

$$t_i = G_i \Big/ (\sum_{j=1}^{n} a_{ij} X_j + y_{D_i}), \quad i = 1, 2, \cdots, n, \tag{3-4}$$

$$s_i = P_i \Big/ \left(\sum_{j=1}^{n} a_{ij}X_j + y_{D_i}\right), \quad i = 1,2,\cdots,n, \tag{3-5}$$

或

$$G = \hat{t}(AX + Y_D), \quad P = \hat{s}(AX + Y_D). \tag{3-6}$$

其中 $\hat{t} = \mathrm{diag}(t_1, t_2, \cdots, t_n)$，$\hat{s} = \mathrm{diag}(s_1, s_2, \cdots, s_n)$.

将(3-6)式代入(3-1)式或(3-2)式，得

$$X = [I - (I - \hat{t} - \hat{s})A]^{-1}[(I - \hat{t} - \hat{s})Y_D + F + O]. \tag{3-7}$$

在直接消耗系数、调入系数、进口系数已知的条件下，由(3-7)式可以计算出地区内最终产品需求量，调出和出口量的变化对地区总产量的影响.

另外，由 $(I - A)^{-1}$ 并不能得出地区完全消耗系数，因其中包含了调入、进口产品的间接消耗.

类似地，由表 3-1、表 3-3、表 3-4、表 3-5 和表 3-6，均可得出相应的数学模型，此处不再详述.

3.2 地区间投入产出模型

地区间投入产出模型是在地区投入产出模型的基础上，为反映地区间广泛的技术经济联系，在综合各地区投入产出表的基础上建立起来的. 地区间投入产出模型种类繁多，如有简单的地区间投入产出模型、基本的地区间投入产出模型和列昂惕夫地区间投入产出模型等. 地区间投入产出表将各地区、各部门产品生产分配使用情况及价值构成在一张表上反映出来，兼有全国投入产出表和地区投入产出表的双重特点和作用. 假设全国由 m 个地区构成，其地区间投入产出模型分述如下.

1. 简单的地区间投入产出模型

(1) 单纯反映地区间调入、调出量的模型. 将 m 个地区投入产出表中调入和调出两项细化，分别列出各表中部门产品分地区的调入、调出量，然后将所有地区投入产出表并列在一起即可.

例如，第 i 地区投入产出表中的调入、调出项可细化为表 3-7 所示.

表 3-7

		调　入	调　出
地　区		$1,\cdots,i-1,i+1,\cdots,m$	$1,\cdots,i-1,i+1,\cdots,m$
部门	1 2 ⋮ n		

(2)地区间列系数模型. 在建立单纯反映地区间调入、调出模型的基础上，通过计算地区间的供应系数和各地区的直接消耗系数等，建立地区间投入产出列系数

模型,并记

$$t_i^{pq}=r_i^{pq}/r_i^q,\quad i=1,2,\cdots,n;\ p,q=1,2,\cdots,m \tag{3-8}$$

表示 q 地区所需 i 部门产品的总量中由 p 地区提供的比重,称 t_i^{pq} 为供应系数.其中 r_i^{pq} 为 p 地区提供 q 地区 i 部门产品的数量,r_i^q 为 q 地区需 i 部门产品的总量.

$$r_i^q=\sum_{j=1}^n a_{ij}^q X_j^q+Y_i^{q0},\quad i=1,2,\cdots,n;\quad q=1,2,\cdots,m, \tag{3-9}$$

其中 a_{ij}^q 为 q 地区内部门产品的直接消耗系数,X_j^q 为 q 地区 j 部门的总产量,Y_i^{q0} 为 q 地区 i 部门最终产品的合计值减去调出量加上调入量.

由(3-8)、(3-9)式可得

$$r_i^{pq}=t_i^{pq}r_i^q=t_i^{pq}\left(\sum_{j=1}^n a_{ij}^q X_j^q+Y_i^{q0}\right),i=1,2,\cdots,n;p,q=1,2,\cdots,m. \tag{3-10}$$

又 $X_i^p=\sum_{q=1}^m r_i^{pq},i=1,2,\cdots,n;p=1,2,\cdots,m$,则

$$X_i^p=\sum_{q=1}^m\left[t_i^{pq}\left(\sum_{j=1}^n a_{ij}^q X_j^q+Y_i^{q0}\right)\right],i=1,2,\cdots,n;p=1,2,\cdots,m. \tag{3-11}$$

(3) 地区间行系数模型

记

$$h_i^{pq}=r_i^{pq}/X_i^p,\ p,q=1,2,\cdots,m;i=1,2,\cdots,n \tag{3-12}$$

表示 p 地区第 i 部门生产的产品分配给 q 地区的比例,称 h_i^{pq} 为地区分配系数.

由此 q 地区从各地区得到的 i 部门产品的供应总量为

$$\sum_{p=1}^m r_i^{pq}=\sum_{p=1}^m h_i^{pq}X_i^p,\quad q=1,2,\cdots,m;\quad i=1,2,\cdots,n. \tag{3-13}$$

q 地区对 i 部门产品的实际需求量 r_i^q 满足(3-9)式.q 地区对 i 部门产品的实际需求量应等于供应总量,即

$$\sum_{p=1}^m h_i^{pq}X_i^p=\sum_{j=1}^n a_{ij}^q X_j^q+Y_i^{q0},\quad q=1,2,\cdots,m;\quad i=1,2,\cdots,n, \tag{3-14}$$

或

$$\sum_{p=1}^m \hat{H}^{pq}X^p=A^qX^q+Y^{q0},q=1,2,\cdots,m, \tag{3-15}$$

其中

$$\hat{H}^{pq}=\mathrm{diag}(h_1^{pq},h_2^{pq},\cdots,h_n^{pq}),X^p=(X_1^p,X_2^p,\cdots,X_n^p)^{\mathrm{T}},$$
$$A^q=(a_{ij}^q)_{n\times n},\ Y^{q0}=(Y_1^{q0},Y_2^{q0},\cdots,Y_n^{q0})^{\mathrm{T}}.$$

(3-15)式可进一步写成

$$HX=\hat{A}X+Y\quad 或\quad X=(H-\hat{A})^{-1}Y\quad 或\quad Y=(H-\hat{A})X, \tag{3-16}$$

其中 $H=(\hat{H}^{pq})_{nm\times nm},X=((X^1)^{\mathrm{T}},(X^2)^{\mathrm{T}},\cdots,(X^m)^{\mathrm{T}})^{\mathrm{T}},\hat{A}=\mathrm{diag}(A^1,A^2,\cdots,A^m),Y=((Y^{10})^{\mathrm{T}},(Y^{20})^{\mathrm{T}},\cdots,(Y^{m0})^{\mathrm{T}})^{\mathrm{T}}$.

2. 基本的地区间投入产出模型

(1) 基本的地区间投入产出表如表 3-8 所示.

表 3-8

投入			产出：中间产品 地区1 部门1 … 部门n	…	中间产品 地区m 部门1 … 部门n	最终产品 地区1	…	地区m	国家m+1	合计	总产出
补偿价值	地区1	部门1	x_{11}^{11} … x_{1n}^{11}	…	x_{11}^{1m} … x_{1n}^{1m}	Y_1^{11}	…	Y_1^{1m}	$Y_1^{1,m+1}$	Y_1^{10}	X_1^1
		⋮	⋮ ⋮		⋮ ⋮	⋮		⋮	⋮	⋮	⋮
		部门n	x_{n1}^{11} … x_{nn}^{11}	…	x_{n1}^{1m} … x_{nn}^{1m}	Y_n^{11}	…	Y_n^{1m}	$Y_n^{1,m+1}$	Y_n^{10}	X_n^1
	…										
	地区m	部门1	x_{11}^{m1} … x_{1n}^{m1}	…	x_{11}^{mm} … x_{1n}^{mm}	Y_1^{m1}	…	Y_1^{mm}	$Y_1^{m,m+1}$	Y_1^{m0}	X_1^m
		⋮	⋮ ⋮		⋮ ⋮	⋮		⋮	⋮	⋮	⋮
		部门n	x_{n1}^{m1} … x_{nn}^{m1}	…	x_{n1}^{mm} … x_{nn}^{mm}	Y_n^{m1}	…	Y_n^{mm}	$Y_n^{m,m+1}$	Y_n^{m0}	X_n^m
	小计										
	固定资产折旧		D_1^1 … D_n^1	…	D_1^m … D_n^m						
新创造价值	劳动报酬		V_1^1 … V_n^1	…	V_1^m … V_n^m						
	社会纯收入		M_1^1 … M_n^1	…	M_1^m … M_n^m						
	总产值		X_1^1 … X_n^1	…	X_1^m … X_n^m						

(2) 基本的地区间投入产出模型各符号的意义参见投入产出表. 由表的横行，有

$$\sum_{q=1}^{m}\sum_{j=1}^{n} x_{ij}^{pq} + Y_i^{p0} = X_i^p, \quad p = 1,2,\cdots,m; \quad i = 1,2,\cdots,n. \tag{3-17}$$

由表的纵列，有

$$\sum_{p=1}^{m}\sum_{i=1}^{n} x_{ij}^{pq} + D_j^q + V_j^q + M_j^q = X_j^q, q = 1,2,\cdots,m; j = 1,2,\cdots,n. \tag{3-18}$$

记 $a_{ij}^{pq} = x_{ij}^{pq}/X_j^q$，表示 q 地区生产单位 j 种产品要消耗 p 地区提供的 i 种产品的数量，$p, q = 1,2,\cdots,m; i,j = 1,2,\cdots,n$. 将 a_{ij}^{pq} 代入(3-17)式得

$$\sum_{q=1}^{m}\sum_{j=1}^{n} a_{ij}^{pq} X_j^q + Y_i^{p0} = X_i^p, \quad p = 1,2,\cdots,m; \quad i = 1,2,\cdots,n, \tag{3-19}$$

或

$$\sum_{q=1}^{m} A^{pq}X^{q} + Y^{p0} = X^{p},\quad p = 1,2,\cdots,m,\tag{3-20}$$

其中 $A^{pq}=(a_{ij}^{pq})_{n\times n}$，$X^{q}=(X_1^{q},X_2^{q},\cdots,X_n^{q})^{\mathrm{T}}$，$Y^{p0}=(Y_1^{p0},Y_2^{p0},\cdots,Y_n^{p0})^{\mathrm{T}}$，$X^{p}=(X_1^{p},X_2^{p},\cdots,X_n^{p})^{\mathrm{T}}$.

记

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{A}^{11} & \cdots & \boldsymbol{A}^{1m} \\ \vdots & & \vdots \\ \boldsymbol{A}^{m1} & \cdots & \boldsymbol{A}^{mm} \end{bmatrix},\ \boldsymbol{Y}=((\boldsymbol{Y}^{10})^{\mathrm{T}},(\boldsymbol{Y}^{20})^{\mathrm{T}},\cdots,(\boldsymbol{Y}^{m0})^{\mathrm{T}})^{\mathrm{T}},$$

$\boldsymbol{X}=((\boldsymbol{X}^{1})^{\mathrm{T}},(\boldsymbol{X}^{2})^{\mathrm{T}},\cdots,(\boldsymbol{X}^{m})^{\mathrm{T}})^{\mathrm{T}}$，则(3-20)式可改写为

$$\boldsymbol{AX}+\boldsymbol{Y}=\boldsymbol{X},\tag{3-21}$$

或

$$\boldsymbol{X}=(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{Y}.\tag{3-22}$$

将 a_{ij}^{pq} 代入(3-18)式得

$$\sum_{p=1}^{m}\sum_{i=1}^{n} a_{ij}^{pq}X_j^{q} + D_j^{q} + V_j^{q} + M_j^{q} = X_j^{q},\ q = 1,2,\cdots,m;\ j = 1,2,\cdots,n,\tag{3-23}$$

或

$$\sum_{p=1}^{m} \widehat{\boldsymbol{C}}^{q}\boldsymbol{X}^{q} + \boldsymbol{D}^{q} + \boldsymbol{V}^{q} + \boldsymbol{M}^{q} = \boldsymbol{X}^{q},\quad q = 1,2,\cdots,m,\tag{3-24}$$

其中 $\boldsymbol{D}^{q},\boldsymbol{V}^{q},\boldsymbol{M}^{q}$ 分别表示 q 地区各部门固定资产折旧的列向量、劳动报酬列向量和社会纯收入列向量. $\widehat{\boldsymbol{C}}^{q}=\mathrm{diag}\left(\sum_{p=1}^{m}\sum_{i=1}^{n}a_{i1}^{pq},\sum_{p=1}^{m}\sum_{i=1}^{n}a_{i2}^{pq},\cdots,\sum_{p=1}^{m}\sum_{i=1}^{n}a_{in}^{pq}\right)$.

记 $\widehat{\boldsymbol{C}}=\mathrm{diag}(\widehat{\boldsymbol{C}}^{1},\widehat{\boldsymbol{C}}^{2},\cdots,\widehat{\boldsymbol{C}}^{m})$，$\boldsymbol{D}=((\boldsymbol{D}^{1})^{\mathrm{T}},(\boldsymbol{D}^{2})^{\mathrm{T}},\cdots,(\boldsymbol{D}^{m})^{\mathrm{T}})^{\mathrm{T}}$，$\boldsymbol{V}=((\boldsymbol{V}^{1})^{\mathrm{T}},(\boldsymbol{V}^{2})^{\mathrm{T}},\cdots,(\boldsymbol{V}^{m})^{\mathrm{T}})^{\mathrm{T}}$，$\boldsymbol{M}=((\boldsymbol{M}^{1})^{\mathrm{T}},(\boldsymbol{M}^{2})^{\mathrm{T}},\cdots,(\boldsymbol{M}^{m})^{\mathrm{T}})^{\mathrm{T}}$，则(3-24)式可进一步改写为

$$(\boldsymbol{I}-\widehat{\boldsymbol{C}})\boldsymbol{X}=\boldsymbol{D}+\boldsymbol{V}+\boldsymbol{M}\ \text{或}\ \boldsymbol{X}=(\boldsymbol{I}-\widehat{\boldsymbol{C}})^{-1}(\boldsymbol{D}+\boldsymbol{V}+\boldsymbol{M}).\tag{3-25}$$

3. 列昂惕夫地区间引力模型

(1)基本假设. 由万有引力定律的基本思想给出地区 p 供给地区 q 的第 i 部门产品数量 X_i^{pq} 的假设如下：

1) X_i^{pq} 与 p 地区第 i 部门产品的数量 X_i^{p0} 成正比；

2) X_i^{pq} 与 q 地区对 i 部门产品的需求量 X_i^{0q} 成正比；

3) X_i^{pq} 与全国第 i 部门产品的总量 X_i^{00} 成反比.

(2)建立模型.

由假设可得

$$X_i^{pq}=Q_i^{pq}X_i^{p0}X_i^{0q}/X_i^{00},\ p,q=1,2,\cdots,m;\ i=1,2,\cdots,n.\tag{3-26}$$

其中 Q_i^{pq} 为常数，由 4 个辅助参数决定，即

$$Q_i^{pq}=(C_i^{p}+K_i^{q})d_i^{pq}\delta_i^{pq},\ p,q=1,2,\cdots,m;\ i=1,2,\cdots,n.\tag{3-27}$$

其中 d_i^{pq} 表示 i 部门产品从 p 地区运至 q 地区所花运费的倒数；δ_i^{pq} 表示 p 地区向 q

地区运送 i 部门产品的可能性，$\delta_i^{pq}=0$ 表示不可能，$\delta_i^{pq}=1$ 表示可能；C_i^p，K_i^q 分别表示 p 地区供应和 q 地区需求 i 部门产品的条件.

4 劳动投入产出模型

劳动投入产出模型是以劳动量或劳动时间表示的投入产出模型，它可以反映劳动量在各个部门的分配和使用情况.

4.1 活劳动消耗的投入产出模型

活劳动消耗投入产出表一般是通过将价值型投入产出表(按不变价格)换算为以劳动消耗量(以人年为单位)表示的投入产出表而得到的.虽然各部门产品的价格与价值存在着程度不同的差异，导致据此计算的劳动消耗系数不能准确符合实际，但在现有条件下，上述做法不失为一种可行方法.活劳动消耗投入产出表的表式及部门分类等基本上同价值型投入产出表，只是采用劳动量(人年、人月、人日等)计量单位.

1.活劳动投入产出表

其表式如表 4-1 所示.

表 4-1 单位：人年

投入			产出									总计
			中间产品				最终产品					
			1	2	…	n	固定资产更新和大修理	积累	消费	合计		
当年物化劳动	产品物化劳动	1	L_{11}	L_{12}	…	L_{1n}	D_1	S_1	W_1	N_1		L_1
		2	L_{21}	L_{22}	…	L_{2n}	D_2	S_2	W_2	N_2		L_2
		⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		⋮
		n	L_{n1}	L_{n2}	…	L_{nn}	D_n	S_n	W_n	N_n		L_n
	固定资产折旧		G_1	G_2	…	G_n	—	—	—	—		—
活劳动	用于必要产品的消耗		L_{v1}	L_{v2}	…	L_{vn}	—	—	—	—		—
	用于剩余产品的消耗		L_{m1}	L_{m2}	…	L_{mn}	—	—	—	—		—
	合计		L_1	L_2	…	L_n	—	—	—	—		—
合计			H_1	H_2	…	H_n	—	—	—	—		—

表中物化劳动部分只包含物化于各部门产品生产中的以人年表示的当年直接劳动消耗量，而未包含全部物化劳动量.

2. 活劳动投入产出表的编制方法

(1) 根据有关的统计和调查资料，确定各部门的年平均工作人员数(以人年计)，设 L_i 表示 i 部门的年平均工作人员数.

(2) 计算劳动消耗系数. 设 t_i 表示各部门单位产品的活劳动消耗系数，则

$$t_i = L_i / X_i, \quad i = 1,2,\cdots,n. \tag{4-1}$$

(3) 设 l_{ij} 表示以人年为单位的 i 部门产品流向 j 部门的数量，则

$$l_{ij} = t_i x_{ij}, \quad i,j = 1,2,\cdots,n. \tag{4-2}$$

由 $X_i = \sum_{j=1}^{n} x_{ij} + Y_i$，得

$$\sum_{j=1}^{n} t_i x_{ij} + t_i Y_i = t_i X_i, \quad i = 1,2,\cdots,n, \tag{4-3}$$

即

$$\sum_{j=1}^{n} l_{ij} + LY_i = L_i, \quad i = 1,2,\cdots,n, \tag{4-4}$$

其中 LY_i 表示以人年为单位的 i 部门产品用于最终产品的数量，即 $LY_i = t_i Y_i$.

(4) 固定资产折旧的劳动量换算. 实际计算中，采用确定当年固定资产更新和大修理的劳动消耗量的方法，以所需的人年数来表示折旧. 记 u_i 表示 i 部门固定资产更新、大修理的劳动消耗系数，则 j 部门折旧的劳动计量公式为

$$G_j = \sum_{i=1}^{n} u_i (d_i K_{ij}), \tag{4-5}$$

其中 K_{ij} 为 j 部门使用 i 类固定资产的价值，d_i 为 i 类固定资产的折旧率.

4.2　完全劳动消耗的投入产出模型

活劳动投入产出表反映的只是当年投入的活劳动在各个部门间的消耗与联系，而不是全部劳动消耗. 所谓**完全劳动消耗**，是指产品生产中所包含的全部物化劳动和活劳动的总量，反映的是产品经过各个不同阶段或部门生产逐步积累起来的劳动消耗的全部，即包含产品的直接劳动消耗和一切间接劳动消耗. 单位产品的完全劳动消耗计算公式为

$$\boldsymbol{T} = \boldsymbol{t}(\boldsymbol{I} - \boldsymbol{A})^{-1} \text{或 } \boldsymbol{T} = \boldsymbol{t}(\boldsymbol{B} + \boldsymbol{I}), \tag{4-6}$$

其中 $\boldsymbol{T} = (T_1, T_2, \cdots, T_n)$ 为各部门单位产品的完全劳动消耗行向量，$\boldsymbol{t} = (t_1, t_2, \cdots, t_n)$ 为各部门单位产品活劳动消耗行向量，$\boldsymbol{A}$，$\boldsymbol{B}$ 分别为产品直接消耗系数和完全消耗系数，$\boldsymbol{I}$ 为 n 阶单位阵.

1. 完全劳动消耗的投入产出表

其表式如表 4-2 所示.

2. 完全劳动消耗的投入产出模型

$$\sum_{j=1}^{n} x_{ij} T_i + Y_i T_i = X_i T_i, \quad i = 1,2,\cdots,n, \tag{4-7}$$

表 4-2 单位:万人年

投入		产出					
		产品部门				最终产品完全劳动量	总产品完全劳动量
		1	2	…	n		
生产部门	1	$x_{11}T_1$	$x_{12}T_2$	…	$x_{1n}T_n$	Y_1T_1	X_1T_1
	2	$x_{21}T_1$	$x_{22}T_2$	…	$x_{2n}T_n$	Y_2T_2	X_2T_2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	$x_{n1}T_1$	$x_{n2}T_2$	…	$x_{nn}T_n$	Y_nT_n	X_nT_n
活劳动		X_1t_1	X_2t_2	…	X_nt_n		
完全劳动消耗量		X_1T_1	X_2T_2	…	X_nT_n		

$$\sum_{i=1}^{n} x_{ij}T_i + X_jt_j = X_jT_j, \quad j = 1,2,\cdots,n, \tag{4-8}$$

$$\sum_{i=1}^{n} x_{ij}T_i + X_jt_j = \sum_{i=1}^{n} x_{ji}T_j + Y_jT_j, \quad j = 1,2,\cdots,n, \tag{4-9}$$

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{n} x_{ij}T_i + X_jt_j\right) = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} x_{ji}T_j + Y_jT_j\right), \tag{4-10}$$

$$\sum_{j=1}^{n} X_jt_j = \sum_{j=1}^{n} Y_jT_j. \tag{4-11}$$

5 动态投入产出模型

特定时间段(如年度)上国民经济的投入产出模型为静态投入产出模型.若要全面连续考察社会再生产过程,必须引入时间变化的概念,即须研究**动态投入产出模型**,这样才能从发展变化中研究社会产品、劳动力、劳动资料与劳动对象间的有机联系及其运动过程.动态投入产出模型主要用于分析、研究生产性积累与社会产品再生产间的内在联系,其理论研究和实际应用受到了各国学者的重视.在静态模型中,投资作为外生变量,与本期生产不发生关系,也无法反映生产性投资与下期生产活动的内在联系.在动态模型中,引进了**资本系数**或**投资系数**矩阵,使外生变量内生化,能够以此连续考察一定时期内投资与再生产的关系.

5.1 建立动态模型要考虑的因素

1. 正确分析生产性积累与非生产性积累

生产性积累是影响生产增长的主要因素,它包含生产性固定资产积累、流动资

产积累及用于更新改造的固定资产积累等,其中生产性固定资产积累是直接用于扩大再生产的,在中国主要采取基本建设的方式进行,受建设项目的投资规模、投资在各年的分配比例、施工技术条件、技术复杂程度等因素的影响,致使基本建设的周期一般较长,大中项目往往不能当年建设、当年竣工投产,常常在若干年后才能形成生产能力,使基本建设的投资和使用之间存在着时间间隔,即"时滞".因此,模型要反映含时间因素在内的生产性固定资产再生产与社会产品再生产间的关系.生产性流动资产的积累为扩大再生产所需的原材料、半成品、成品等各种储备,即为劳动对象的积累.生产性固定资产更新改造一般认为是固定资产的简单再生产,但固定资产折旧与更新不是同时进行的,往往需经过一段时间折旧之后才更新,更新常伴随着科技的进步,且固定资产折旧基金常与积累基金混合使用,都作为固定资产的投资,因此,固定资产的更新改造就同时发挥着简单再生产和扩大再生产两方面的作用.非生产性积累主要用于下期消费,不能直接用于生产.

2. 适当确定动态模型的考察期

动态模型研究生产性固定资产再生产与社会产品再生产间的关系,如果考察时期过短则难以清楚全面地揭示和认识这种关系,考察时期过长又存在着资料搜集困难、结果失真等缺点,因此其考察时间必须适中.

5.2　几种动态投入产出模型

5.2.1　列昂惕夫动态投入产出模型

1953 年列昂惕夫在《美国经济结构研究》中提出了用微分方程表示的动态投入产出模型.模型分为封闭式与开启式两种.其中封闭式模型将居民也作为一个部门,其产出是向国民经济各部门提供的劳务,投入就是居民的各种消费;而开启式模型是把包括居民消费在内的一些因素放在最终产品中,作为模型的外生变量.

1. 封闭式动态模型

$$X_i^{(t)} - \sum_{j=1}^{n} a_{ij} X_j^{(t)} - \sum_{j=1}^{n} b_{ij} \frac{\mathrm{d}X_j^{(t)}}{\mathrm{d}t} = 0, \quad i = 1,2,\cdots,n. \tag{5-1}$$

其中 $b_{ij} = F_{ij}/\Delta X_j$ 表示 j 部门增加单位产品所需占用 i 部门的产品作为生产资本的数量,包括所需的机器设备、原材料、备件等,称 $b_{ij}(i,j = 1,2,\cdots,n)$ 为**资本系数**. $X_i^{(t)}$ 是 t 年 i 部门的总产量.

2. 开启式动态模型

列昂惕夫考虑到客观经济中有政府活动和进出口等因素存在,不可能呈封闭状态,所以又提出了开启式模型.即

$$X_i^{(t)} - \sum_{j=1}^{n} a_{ij} X_j^{(t)} - \sum_{j=1}^{n} b_{ij} \frac{\mathrm{d}X_j^{(t)}}{\mathrm{d}t} = Y_i^{(t)}, \quad i = 1,2,\cdots,n, \tag{5-2}$$

其中 $Y_i^{(t)}$ 是 t 年 i 部门的最终净产品量(不含投资).

3. 差分形式的动态模型

考虑到实际统计资料的离散性,1965 年列昂惕夫又提出了以差分形式表示的

动态模型.

(1) 基本假设.

1)各部门新增的生产能力都能得到充分利用;

2)模型考察期内 $a_{ij}, b_{ij}(i,j=1,2,\cdots,n)$不变;

3)第 t 年生产性资本的增加会引起第 $t+1$ 年生产能力的增加,即时滞为 1 年.

(2) 差分模型.可将微分方程形式的模型改造得到差分方程形式的模型,例如可将开启式模型改写为

$$X_i^{(t)} - \sum_{j=1}^{n} a_{ij} X_j^{(t)} - \sum_{j=1}^{n} b_{ij}(X_j^{(t+1)} - X_j^{(t)}) = Y_i^{(t)}, \quad i = 1,2,\cdots,n. \tag{5-3}$$

或

$$X^{(t)} - AX^{(t)} - B[X^{(t+1)} - X^{(t)}] = Y^{(t)}, \tag{5-4}$$

$$(I - A + B)X^{(t)} - BX^{(t+1)} = Y^{(t)}, \tag{5-5}$$

$$X^{(t)} = (I - A + B)^{-1}[Y^{(t)} + BX^{(t+1)}], \tag{5-6}$$

其中 $X^{(t)} = (X_1^{(t)}, X_2^{(t)}, \cdots, X_n^{(t)})^{\mathrm{T}}, A = (a_{ij})_{n\times n}, B = (b_{ij})_{n\times n}, Y^{(t)} = (Y_1^{(t)}, Y_2^{(t)}, \cdots, Y_n^{(t)})^{\mathrm{T}}$.

4. 考虑时变的动态模型

1970 年,列昂惕夫在差分方程动态模型的基础上提出了考虑时变的动态模型,即 a_{ij}和 b_{ij}随年份变化而变化.模型为

$$X^{(t)} - A^{(t)}X^{(t)} - B^{(t+1)}[X^{(t+1)} - X^{(t)}] = Y^{(t)}, \tag{5-7}$$

由此得

$$[I - A^{(t)} + B^{(t+1)}]X^{(t)} - B^{(t+1)}X^{(t+1)} = Y^{(t)}, \tag{5-8}$$

或

$$G^{(t)}X^{(t)} - B^{(t+1)}X^{(t+1)} = Y^{(t)}. \tag{5-9}$$

其中 $G^{(t)} = I - A^{(t)} + B^{(t+1)}$, $Y^{(t)}$表示第 t 年的最终净产品量(不含投资).

用(5-9)式论证经济的实际发展过程时,若往前考察 m 年的情况,可采用下面的递推方法.

记基年为 0 年, $-t$ 年表示往前 t 年,假设 $Y^{(0)}$为基年包括投资的最终产品量,则递推过程为

$$\begin{cases} X^{(0)} = [I - A^{(0)}]^{-1}Y^{(0)}; & (5\text{-}10) \\ X^{(-t)} = [G^{(-t)}]^{-1}[Y^{(-t)} + B^{(-t+1)}X^{(-t+1)}], \quad t = 1,2,\cdots,m. & (5\text{-}11) \end{cases}$$

类似地,用上述方法可以考察基年至计划期(第 m 年)的经济变化过程,但尚需事先对 $A^{(t)}, B^{(t+1)}, Y^{(t)}$等进行预测和估计.

5.2.2 其他动态模型

1. 考虑生产性的固定资产占用模型

$$X_i^{(t)} - \sum_{j=1}^{n} a_{ij}X_j^{(t)} - \sum_{j=1}^{n}[F_{ij}^{(t+1)} - F_{ij}^{(t)}] = Y_i^{(t)}, \quad i = 1,2,\cdots,n, \tag{5-12}$$

其中 $F_{ij}^{(t)}$为第 t 年 j 部门生产时所需占用 i 种生产性固定资产的数量, $Y_i^{(t)}$ 为第 t

年的最终净产品(即最终产品中扣除生产性固定资产积累的余额).

(1)若引入**生产性固定资产的占用系数** $f_{ij}, f_{ij}=F_{ij}^{(t)}/X_j^{(t)}$,则(5-12)式可改为

$$X_i^{(t)}-\sum_{j=1}^{n}a_{ij}X_j^{(t)}-\sum_{j=1}^{n}f_{ij}[X_j^{(t+1)}-X_j^{(t)}]=Y_i^{(t)},\quad i=1,2,\cdots,n,\tag{5-13}$$

(2)若 $F_{ij}^{(t)}$ 为第 t 年年初 j 部门占用 i 种固定资产的数量,令 $\overline{f}_{ij}=F_{ij}^{(t+1)}/X_j^{(t)}$,则(5-12)式可改写为

$$X_i^{(t)}-\sum_{j=1}^{n}a_{ij}X_j^{(t)}-\sum_{j=1}^{n}[\overline{f}_{ij}X_j^{(t)}-F_{ij}^{(t)}]=Y_i^{(t)},\quad i=1,2,\cdots,n,\tag{5-14}$$

或

$$[\boldsymbol{I}-\boldsymbol{A}-\overline{\boldsymbol{f}}]\boldsymbol{X}^{(t)}=\boldsymbol{Y}^{(t)}-\boldsymbol{F}^{(t)},\tag{5-15}$$

$$\boldsymbol{X}^{(t)}=[\boldsymbol{I}-\boldsymbol{A}-\overline{\boldsymbol{f}}]^{-1}[\boldsymbol{Y}^{(t)}-\boldsymbol{F}^{(t)}],\tag{5-16}$$

其中

$$\boldsymbol{F}^{(t)}=[F_1^{(t)},F_2^{(t)},\cdots,F_n^{(t)}]^{\mathrm{T}},\quad F_i^{(t)}=\sum_{j=1}^{n}F_{ij}^{(t)},$$

$$\boldsymbol{Y}^{(t)}=(Y_1^{(t)},Y_2^{(t)},\cdots,Y_n^{(t)})^{\mathrm{T}},\quad \overline{\boldsymbol{f}}=(\overline{f}_{ij})_{n\times n}.$$

2. 考虑科技进步和增量的固定资产占用模型

记

$$\overline{a}_{ij}=\Delta x_{ij}/\Delta X_j=(x_{ij}^{(t)}-x_{ij}^{(t-1)})/(X_j^{(t)}-X_j^{(t-1)})\tag{5-17}$$

表示每增加一个单位 j 部门产品需直接增加消耗 i 部门产品的数量,称 $\overline{a}_{ij}$ 为**增量直接消耗系数**.在科技进步条件下,其值一般低于第 $t-1$ 年的直接消耗系数.记 $\widehat{f}_{ij}=\Delta F_{ij}/\Delta X_j$ 表示每增加一个单位 j 部门产品需增加占用 i 种固定资产的数量,称 $\widehat{f}_{ij}$ 为**增量固定资产占用系数**.

将(5-12)式改写为

$$X_i^{(t)}-\sum_{j=1}^{n}\overline{a}_{ij}X_j^{(t)}-\sum_{j=1}^{n}\widehat{f}_{ij}(X_j^{(t)}-X_j^{(t-1)})=Y_i^{(t)},\quad i=1,2,\cdots,n,\tag{5-18}$$

或

$$\boldsymbol{X}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{X}^{(t)}-\widehat{\boldsymbol{f}}\boldsymbol{X}^{(t)}=\boldsymbol{Y}^{(t)}-\widehat{\boldsymbol{f}}\boldsymbol{X}^{(t-1)},\tag{5-19}$$

$$\boldsymbol{X}^{(t)}=[\boldsymbol{I}-\overline{\boldsymbol{A}}-\widehat{\boldsymbol{f}}]^{-1}(\boldsymbol{Y}^{(t)}-\widehat{\boldsymbol{f}}\boldsymbol{X}^{(t-1)}),\tag{5-20}$$

其中 $\overline{\boldsymbol{A}}=(\overline{a}_{ij})_{n\times n}$, $\widehat{\boldsymbol{f}}=(\widehat{f}_{ij})_{n\times n}$.

3. 半动态投入产出模型

该模型主要用来确定计划期最终年的各种指标,它是通过利用产品与固定资产静态投入产出模型的资料而建立起来的,除通常考虑的因素外,还考虑诸如生产储备的增加、非生产性固定资产积累的增加、固定资产的磨损等因素,且通过相应参数来反映固定资产积累的变化与产品产量的变化以及变化间的联系,从而简化了模型的求解和运算.

假设国民经济的部门分成两大类,一类是生产劳动对象与消费品的部门,记为 $1,2,\cdots,l$ 部门;另一类是生产建筑产品与机器设备等固定资产部门,记为 $l+1$, $l+2,\cdots,n$ 部门.

将产品静态投入产出表稍微改造,以主要反映第一类各部门与所有部门间的投入产出关系,其表式如表 5-1 所示;构造固定资产静态投入产出表,以主要反映第二类各部门与所有部门间的投入产出关系,其表式如表 5-2 所示.

由表 5-1 可得.

$$X_i^{(t)} - \sum_{j=1}^{n} x_{ij}^{(t)} - h_i^{(t)} = W_i^{(t)} + e_i^{(t)}, \quad i = 1,2,\cdots,l, \tag{5-21}$$

其中,$h_i^{(t)}$, $W_i^{(t)}$ 和 $e_i^{(t)}$ 分别为 t 年 i 部门储备的增加、t 年 i 部门产品用于个人及社会消费的数量以及 t 年 i 部门产品流入净额.

(5-21)式可改写为含直接消耗系数 a_{ij} 的方程

$$X_i^{(t)} - \sum_{j=1}^{n} a_{ij}X_j^{(t)} - h_i^{(t)} = W_i^{(t)} + e_i^{(t)}, \quad i = 1,2,\cdots,l. \tag{5-22}$$

由表 5-2 可得

$$F_i^{(t)} - \sum_{j=1}^{n} F_{ij}^{(t)} - h_i^{(t)} - K_i^{(2)(t)} = G_i^{(t)} + K_i^{(1)(t)} + e_i^{(t)}, \quad i = l+1,\cdots,n. \tag{5-23}$$

其中 $F_i^{(t)}$ 为 t 年末 i 部门固定资产占有量.

(5-23)式可改写为含固定资产占用系数 f_{ij} 的方程

$$F_i^{(t)} - \sum_{j=1}^{n} f_{ij}X_j^{(t)} - h_i^{(t)} - K_i^{(2)(t)} = G_i^{(t)} + K_i^{(1)(t)} + e_i^{(t)}, \quad i = l+1,\cdots,n. \tag{5-24}$$

假设规划期内每年生产性固定资产积累的平均增长率为 δ,则有

$$K_i^{(1)(t)} = K_i^{(1)(0)}(1+\delta)^t, \tag{5-25}$$

其中 $K_i^{(1)(0)}$ 为基年第 i 种固定资产生产性积累的数量. 又假设已知固定资产的磨损系数、各种产品与固定资产储备的增长率、社会消费与个人消费的增长率、单位基建投资能使生产性固定资产增加的系数等,则由(5-21)式(或(5-22)式)、(5-23)式(或(5-24)式)和(5-25)式可求出规划期最终年各部门产品的生产总量、各部门固定资产的总量及各种固定资产的积累量. 因此,半动态投入产出模型主要用于确定规划期最终年的各种指标.

4. 综合动态模型

半动态模型中有许多平均变化率的假设,有时与实际有一定偏差,为了能够切实反映生产性固定资产积累与社会产品生产的逐年变化,这里给出了综合动态模型. 综合动态模型要考虑的因素很多,形成一个由产品动态平衡方程、基建投资方程、劳动力平衡方程和最终净产品方程等四个方程组构成的方程体系. 求解该方程体系便可得到规划期各年度的生产性投资与各部门产量. 各方程的具体内容如下.

表

投入		产							
		中间产品							
		1	2	…	l	$l+1$	…	n	合计
物资消耗	部门 1	x_{11}	x_{12}	…	x_{1l}	$x_{1,l+1}$	…	x_{1n}	$\sum_{j=1}^{n} x_{1j}$
	2	x_{21}	x_{22}	…	x_{2l}	$x_{2,l+1}$	…	x_{2n}	$\sum_{j=1}^{n} x_{2j}$
	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
	l	x_{l1}	x_{l2}	…	x_{ll}	$x_{l,l+1}$	…	x_{ln}	$\sum_{j=1}^{n} x_{lj}$
	$l+1$								
	⋮								
	n								
	小计								
	固定资产折旧	D_1	D_2	…	D_l	D_{l+1}	…	D_n	D
	合计	C_1	C_2	…	C_l	C_{l+1}	…	C_n	C
新创价值	劳动报酬	V_1	V_2	…	V_l	V_{l+1}	…	V_n	V
	社会纯收入	m_1	m_2	…	m_l	m_{l+1}	…	m_n	M
	合计	N_1	N_2	…	N_l	N_{l+1}	…	N_n	N
总计		X_1	X_2	…	X_l	X_{l+1}	…	X_n	X

表

固定资产的生产部门	基期固定资产总量	固定资产的					
		本期生产中占用的固定资产					
		部门 1	部门 2	…	部门 l	部门 $l+1$	…
部门 $l+1$	$F_{l+1}^{(0)}$	$F_{l+1,1}$	$F_{l+1,2}$	…	$F_{l+1,l}$	$F_{l+1,l+1}$	…
$l+2$	$F_{l+2}^{(0)}$	$F_{l+2,1}$	$F_{l+2,2}$	…	$F_{l+2,l}$	$F_{l+2,l+1}$	…
⋮	⋮	⋮	⋮		⋮	⋮	
n	$F_n^{(0)}$	F_{n1}	F_{n2}	…	F_{nl}	$F_{n,l+1}$	…
合计	$F^{(0)}$	$\sum_{i=l+1}^{n} F_{i1}$	$\sum_{i=l+1}^{n} F_{i2}$	…	$\sum_{i=l+1}^{n} F_{il}$	$\sum_{i=l+1}^{n} F_{i,l+1}$	…

5-1

单位:亿元

出							
最终产品							总计
固定资产更新与大修	个人消费与社会消费	储备的增加	投资		进出口平衡	合计	
			生产性	非生产性			
—	W_1	h_1	—	—	e_1	y_1	X_1
—	W_2	h_2	—	—	e_1	y_2	X_2
—	⋮	⋮	—	—	⋮	⋮	⋮
—	W_l	h_l	—	—	e_l	y_l	X_l
G_{l+1}	—	h_{l+1}	$K_{l+1}^{(1)}$	$K_{l+1}^{(2)}$	e_{l+1}	y_{l+1}	X_{l+1}
⋮	—	⋮	⋮	⋮	⋮	⋮	⋮
G_n	—	h_n	$K_n^{(1)}$	$K_n^{(2)}$	e_n	y_n	X_n
G	W	H	$K^{(1)}$	$K^{(2)}$	E	Y	X

5-2

使用部门								
		固定资产的更新、积累及其他						总计
部门 n	合计	更新	增加储备	生产性积累	非生产性积累	进出口平衡	合计	
$F_{l+1,n}$	$\sum_{j=1}^{n} F_{l+1,j}$	G_{l+1}	h_{l+1}	$K_{l+1}^{(1)}$	$K_{l+1}^{(2)}$	e_{l+1}	Y_{l+1}	F_{l+1}
$F_{l+2,n}$	$\sum_{j=1}^{n} F_{l+2,j}$	G_{l+2}	h_{l+2}	$K_{l+2}^{(1)}$	$K_{l+2}^{(2)}$	e_{l+2}	Y_{l+2}	F_{l+2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
F_{nn}	$\sum_{j=1}^{n} F_{nj}$	G_n	h_n	$K_n^{(1)}$	$K_n^{(2)}$	e_n	Y_n	F_n
$\sum_{i=l+1}^{n} F_{in}$		G	H'	$K^{(1)}$	$K^{(2)}$	E'	Y'	F

表

生　　产		分配									
		中间产品与生产性									
		中　　间　　产　　品									
		1	2	…	l	$l+1$	…	n	合　计	1	2
物资消耗	1	x_{11}	x_{12}	…	x_{1l}	$x_{1,l+1}$	…	x_{1n}	$\sum_{j=1}^{n} x_{1j}$		
	2	x_{21}	x_{22}	…	x_{2l}	$x_{2,l+1}$	…	x_{2n}	$\sum_{j=1}^{n} x_{2j}$		
	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		
	l	x_{l1}	x_{l2}	…	x_{ll}	$x_{l,l+1}$	…	x_{ln}	$\sum_{j=1}^{n} x_{lj}$		
	$l+1$									$K_{l+1,1}^{(1)}$	$K_{l+1,2}^{(1)}$
	⋮									⋮	⋮
	n									$K_{n1}^{(1)}$	$K_{n2}^{(1)}$
	小　计	$\sum_{i=1}^{l} x_{i1}$	$\sum_{i=1}^{l} x_{i2}$	…	$\sum_{i=1}^{l} x_{il}$	$\sum_{i=1}^{l} x_{i,l+1}$	…	$\sum_{i=1}^{l} x_{in}$	$\sum_{i=1}^{l}\sum_{j=1}^{n} x_{ij}$	$K_1^{(1)}$	$K_2^{(1)}$
	固定资产折旧	D_1	D_2	…	D_l	D_{l+1}	…	D_n	D		
	合　计	C_1	C_2	…	C_l	C_{l+1}	…	C_n	C		
新创造价值	劳动报酬	V_1	V_2	…	V_l	V_{l+1}	…	V_n	V		
	社会纯收入	m_1	m_2	…	m_l	m_{l+1}	…	m_n	m		
	合　计	N_1	N_2	…	N_l	N_{l+1}	…	N_n	N		
总　　计		X_1	X_2	…	X_l	X_{l+1}	…	X_n	X		

①　表中的生产性固定资产积累，实际上包括了固定资产的更新，确切地说是固定资产投资．

5-3

使用

固定资产积累						最终净产品		总计
生产性固定资产积累[①] $K(1)$						上期最终净产品	本期最终净产品	
…	l	$l+1$	…	n	合计			
						$\overline{Y}_1^{(0)}$	$\overline{Y}_1$	X_1
						$\overline{Y}_2^{(0)}$	$\overline{Y}_2$	X_2
						⋮	⋮	⋮
						$\overline{Y}_l^{(0)}$	$\overline{Y}_l$	X_l
…	$K_{l+1,l}^{(1)}$	$K_{l+1,l+1}^{(1)}$	…	$K_{l+1,n}^{(1)}$	$\sum_{j=1}^{n} K_{l+1,j}^{(1)}$	$\overline{Y}_{l+1}^{(0)}$	$\overline{Y}_{l+1}$	X_{l+1}
	⋮	⋮		⋮	⋮	⋮	⋮	⋮
…	$K_{nl}^{(1)}$	$K_{n,l+1}^{(1)}$	…	$K_{n,n}^{(1)}$	$\sum_{j=1}^{n} K_{nj}^{(1)}$	$\overline{Y}_n^{(0)}$	$\overline{Y}_n$	X_n
…	$K_l^{(1)}$	$K_{l+1}^{(1)}$	…	$K_n^{(1)}$	$K^{(1)}$	$\overline{Y}^{(0)}$	$\overline{Y}$	X

(1) 产品动态平衡方程.

产品动态平衡表如表 5-3 所示,在表 5-3 中,将生产性固定资产积累从最终产品栏移至中间产品栏,使它变为内生变量;将不包括生产性固定资产积累的最终产品(即最终净产品)作为外生变量.与半动态模型类似,在表 5-3 中,国民经济也相应分成两大类部门,其平衡方程为

$$X_i^{(t)} = \sum_{j=1}^{n} x_{ij}^{(t)} + \sum_{j=1}^{n} K_{ij}^{(1)(t)} + \overline{Y}_i^{(0)} + \beta_i\left(\sum_{i=1}^{n}\overline{Y}_i^{(t)} - \sum_{i=1}^{n}\overline{Y}_i^{(0)}\right), \qquad (5\text{-}26)$$

用直接消耗系数 $a_{ij} = x_{ij}^{(t)}/X_j^{(t)}$ 及生产性投资构成系数 $K_{ij} = K_{ij}^{(1)(t)}/K_j^{(1)(t)}$, $i = l+1,\cdots,n$; $j = 1,2,\cdots,n$, 代入方程(5-26),得

$$X_i^{(t)} = \sum_{j=1}^{n} a_{ij}X_j^{(t)} + \sum_{j=1}^{n} K_{ij}K_j^{(1)(t)} + \overline{Y}_i^{(0)} + \beta_i\left(\sum_{i=1}^{n}\overline{Y}_i^{(t)} - \sum_{i=1}^{n}\overline{Y}_i^{(0)}\right), \quad (5\text{-}27)$$

其中 K_{ij} 表示 j 部门一个单位生产性固定资产积累(投资)所需第 i 种固定资产的数量, $\sum_{i=1}^{n}\overline{Y}_i^{(t)} - \sum_{i=1}^{n}\overline{Y}_i^{(0)}$ 为最终净产品的增量, β_i 为 i 部门最终净产品增量占总增量的比重, $K_j^{(1)(t)}(j=1,2,\cdots,n)$ 要通过基建投资方程计算.

(2) 基建投资方程.建立基建投资方程,需考虑下列几个因素:一是产品生产中对固定资产的需要,这可通过利用固定资产静态模型中的固定资产占用系数 f_{ij} 来建立这一关系;二是固定资产更新的需要,这可通过计算固定资产的报废率来得到有关数据;三是固定资产时滞的变化,这可通过计算未完工程增长额占新增固定资产的比重来近似反映;四是其他基建投资的数量,这可通过计算其他投资占总投资的比重得到.

投资方程为

$$K_j^{(1)(t)} = \left(\frac{\overline{f}_jX_j^{(t)} - F_j^{(t-1)}}{g_j} + W_j^{(t)}F_j^{(t-1)}\right)\frac{1+\psi_j^{(t)}}{1-\alpha_j^{(t)}}, \qquad (5\text{-}28)$$

其中 $\overline{f}_j$ 为 j 部门的固定资产占用系数, $\overline{f}_j = \sum_{i=l+1}^{n} f_{ij}$, $F_j^{(t-1)}$ 为第 $t-1$ 年年末 j 部门生产性固定资产总量, g_j 为第 t 年末固定资产转换成年平均量的系数,即为固定资产全年增量与年平均增量的比率, $W_j^{(t)}$ 为第 t 年 j 部门固定资产的报废率, $\psi_j^{(t)}$ 为第 t 年 j 部门未完工程增长额与新增固定资产的比率, $\alpha_j^{(t)}$ 为第 t 年其他基建投资占总投资的比重.

(3) 劳动力平衡方程.该方程是通过劳动消耗系数 t_j 建立的, t_j 表示生产单位 j 部门产品需直接消耗劳动力的数量(以价值表现),平衡方程为

$$L_j^{(t)} = t_jX_j^{(t)}, \quad j = 1,2,\cdots,n. \qquad (5\text{-}29)$$

(4) 最终净产品方程.通过基建投资占最终产品的比重 $r^{(t)}$ 来建立最终净产品与基建投资额之间的联系,方程为

$$\sum_{i=1}^{n}\overline{Y}_i^{(t)} = \sum_{j=1}^{n} K_j^{(1)(t)}\left(\frac{1}{r(t)} - 1\right), \qquad (5\text{-}30)$$

其中，$\sum_{i=1}^{n}\overline{Y}_i^{(t)}$ 为第 t 年国民经济各部门最终净产品总额，

$$r^{(t)}=\sum_{j=1}^{n}K_j^{(1)(t)}\Big/\sum_{j=1}^{n}\left[K_j^{(1)(t)}+\overline{Y}_j^{(t)}\right].$$

综上分析，综合动态模型将上述四组方程联立起来，配合使用规划和统计工作中的其他资料，从动态变化中考察国民经济，通过它能分析、研究与制订规划期内国民经济各部门的产量、固定资产积累量、劳动力需要量及最终净产品数量等指标.

6 投入产出优化模型

编制投入产出表的目的，在于对国民经济各部门的经济运作状况及部门间的经济技术联系等情况进行考察，为研究制定计划、进行经济预测等提供基础资料，为优化产业结构、提高经济效益和社会效益、最优配置人力资源以及自然资源和财力资源、保护环境资源等提供决策依据.最后一条正是**投入产出优化模型**所要完成的任务，也是投入产出方法的研究和应用相当活跃的一个方面.

投入产出优化模型是在投入产出静态模型、动态模型与线性规划、非线性规划或多目标规划等相结合的基础上形成的.其中投入产出静态模型与线性规划相结合所形成的线性规划模型，是研究较早的投入产出优化模型.

6.1 投入产出线性规划模型

6.1.1 最简单的两种模型

1.模型Ⅰ

设 $Y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 为社会在一定时期内对最终产品的需要，$X=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为各部门在该时期的产量，$A=(a_{ij})_{n\times n}$为直接消耗系数，$V=(V_1,V_2,\cdots,V_n)^{\mathrm{T}}$ 为直接劳动消耗系数，则各部门的产量在扣除生产中消耗的中间产品后，应大于或等于社会对最终产品的需要量，即

$$(I-A)X\geqslant Y.\tag{6-1}$$

若要求生产各部门产品的劳动消耗量最小，则目标函数为

$$\min V^{\mathrm{T}}X,\tag{6-2}$$

考虑产量的实际情况，有

$$X_i\geqslant 0,\quad i=1,2,\cdots,n,\tag{6-3}$$

综合(6-1)式～(6-3)式，确定各部门产量的使劳动消耗量最小的线性规划模型可表述为

$$\begin{cases} \min & V^{\mathrm{T}}X; \\ \text{s.t.} & (I-A)X \geqslant Y, \\ & X \geqslant 0, \end{cases} \tag{6-4}$$

相应的最优解 X^*,即各部门最优的产量为 $X^*=(I-A)^{-1}Y$,最小的劳动消耗为 $V^{\mathrm{T}}[(I-A)^{-1}Y]$.

2. 模型Ⅱ

设 $P=(P_1,P_2,\cdots,P_n)^{\mathrm{T}}$,其中 P_i 为 i 部门生产过程中单位产品的完全劳动消耗量,$P_i \geqslant 0, i=1,2,\cdots,n$.由投入产出表的纵列知,提供给国民经济中各部门单位产品的中间消耗所对应的劳动力消耗与劳动力直接消耗之和应不小于单位产品的完全劳动消耗,即

$$A^{\mathrm{T}}P+V \geqslant P, \tag{6-5}$$

或

$$(I-A^{\mathrm{T}})P \leqslant V. \tag{6-6}$$

若使价值形态的国民收入(最终产品)最大,则目标函数应为

$$\max Y^{\mathrm{T}}P, \tag{6-7}$$

因此,相应的线性规划模型可表述为

$$\begin{cases} \max & V^{\mathrm{T}}P; \\ \text{s.t.} & (I-A^{\mathrm{T}})P \leqslant V, P \geqslant 0, \end{cases} \tag{6-8}$$

相应的最优解 P^*,即各部门的完全劳动消耗量为 $P^*=(I-A^{\mathrm{T}})^{-1}V$,国民收入的最大值为 $Y^{\mathrm{T}}[(I-A^{\mathrm{T}})^{-1}V]$.

6.1.2　其他线性规划模型

(1) 在投入产出静态模型中,除已知的表示部门间经济技术联系的平衡方程约束外,一般还应考虑劳动力可能变化的程度,自然资源尤其是某些稀缺物资的供应量,各部门资金占用量等限制,使国内生产总值(或社会总产值)最大,相应的模型则为

$$\begin{cases} \max & \sum_{i=1}^{n} Y_i (\text{或} \sum_{i=1}^{n} X_i); \\ \text{s.t.} & (I-A)X=Y; \\ & B^{\mathrm{T}}X \leqslant L; \\ & C^{\mathrm{T}}X \leqslant K; \\ & D^{\mathrm{T}}X \leqslant W; \\ & E^{\mathrm{T}}X \leqslant I; \\ & 0 \leqslant X \leqslant F, Y \geqslant 0. \end{cases} \tag{6-9}$$

其中 $B=(b_1,b_2,\cdots,b_n)^{\mathrm{T}}$, b_i 为 i 部门单位产品消耗劳动力的系数;$C=(c_1,c_2,\cdots,c_n)^{\mathrm{T}}$, c_i 为 i 部门生产单位产品消耗某种自然资源的系数;$D=(d_1,d_2,\cdots,d_n)^{\mathrm{T}}$, d_i 为 i 部门生产单位产品消耗某种稀缺资源的系数;$E=(e_1,e_2,\cdots,e_n)^{\mathrm{T}}$, e_i 为 i 部门生产单位产品占用资金的系数;$F=(f_1,f_2,\cdots,f_n)^{\mathrm{T}}$, f_i 为 i 部门生产能力的限制;

L,K,W,I 分别是社会所能提供的劳动力数量、某种自然资源的数量、某种稀缺资源的数量及资金的数量.若自然资源或稀缺资源的限制种类不止一种,应分别列出相应的约束条件.

(2) 考虑不同地区的生产力水平与生产状况,确定生产在地区分布上的最优方案.

1) 确定某部门(或某项产品)的生产地区.若该项产品由 n 个地区生产,每个地区有 p 种生产技术生产该产品,且有 m 个地区需要该产品,设 $x_{ij}^{(k)}$ 表示地区 i(产地)供给地区 j(需求地)的以第 k 种技术生产的该项产品的量,$X_i^{(k)}$ 表示地区 i 可以提供的以第 k 种技术生产的该产品最大量,Z_j 表示地区 j 需要该产品的量,$C_{ij}^{(k)}$ 表示地区 i 供给地区 j 以第 k 种技术生产的单位产品的成本,由此可以建立下列使总成本最小的线性规划模型.

目标函数

$$\min\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p}x_{ij}^{(k)}C_{ij}^{(k)}, \tag{6-10}$$

约束条件中除 $x_{ij}^{(k)}\geqslant 0, i=1,2,\cdots,n; j=1,2,\cdots,m; k=1,2,\cdots,p$ 外,其他约束条件需根据供应量和需求量的关系确定.

①若 $\sum_{i=1}^{n}\sum_{k=1}^{p}x_i^{(k)} > \sum_{j=1}^{m}Z_j$,即总供应量大于总需求量,则约束为

$$\begin{cases}\sum_{j=1}^{m}x_{ij}^{(k)} \leqslant X_i^{(k)}, & i=1,2,\cdots,n,k=1,2,\cdots,p;\\ \sum_{i=1}^{n}\sum_{k=1}^{p}x_{ij}^{(k)} = Z_j, & j=1,2,\cdots,m.\end{cases} \tag{6-11}$$

②若 $\sum_{i=1}^{n}\sum_{k=1}^{p}X_i^{(k)} = \sum_{j=1}^{m}Z_j$,即总供应量等于总需求量,则约束为

$$\begin{cases}\sum_{j=1}^{m}x_{ij}^{(k)} = X_i^{(k)}, & i=1,2,\cdots,n,k=1,2,\cdots,p;\\ \sum_{i=1}^{n}\sum_{k=1}^{p}x_{ij}^{(k)} = Z_j, & j=1,2,\cdots,m.\end{cases} \tag{6-12}$$

③若 $\sum_{i=1}^{n}\sum_{k=1}^{p}X_i^{(k)} < \sum_{j=1}^{m}Z_j$,即总供应量小于总需求量,则约束为

$$\begin{cases}\sum_{j=1}^{m}x_{ij}^{(k)} = X_i^{(k)}, & i=1,2,\cdots,n,k=1,2,\cdots,p;\\ \sum_{i=1}^{n}\sum_{k=1}^{p}x_{ij}^{(k)} \leqslant Z_j, & j=1,2,\cdots,m.\end{cases} \tag{6-13}$$

上述模型实质上为线性规划中的运输问题.

2) 确定各部门(或各类产品)的生产地区.在已知地区间投入产出模型(参见第3章)的基础上,欲使各地区总劳动消耗量最小,相应模型应为

$$
\begin{cases}
\min \sum_{q=1}^{m} (V^q)^{\mathrm{T}} X^q; \\
\text{s.t. } X_i^p - \sum_{q=1}^{m} \sum_{j=1}^{n} a_{ij}^{pq} X_j^q = Y_i^{p0}, \quad p = 1,2,\cdots,m,\ i = 1,2,\cdots,n; \\
\qquad X_i^p \geqslant 0,\ Y_i^{p0} \geqslant 0,\ p = 1,2,\cdots,m,\ i = 1,2,\cdots,n.
\end{cases}
\tag{6-14}
$$

符号具体意义可参见表3-8及(3-19)式.

6.2　单目标动态投入产出优化模型

建立单目标动态投入产出优化模型的目的,是在投入产出动态模型的基础上,在自然资源、人力资源和资金约束等条件下,确定各部门在规划期内各年度的产值,使规划期内国内生产总值累计最大或劳动力消耗最少或能耗最低等.比如,建立使规划期内国内生产总值累计最大的模型可表述为

$$
\begin{cases}
\max \quad \sum_{t=1}^{T} \sum_{i=1}^{n} f_i(t); \\
\text{s.t.} \quad X_i^{(t)} = \sum_{j=1}^{n} a_{ij}^{(t)} X_j^{(t)} + f_i^{(t)};\ i = 1,2,\cdots,n;\ t = 1,2,\cdots,T; \\
\qquad [\boldsymbol{B}^{(t)}]^{\mathrm{T}} \boldsymbol{X}^{(t)} \leqslant L^{(t)}; \\
\qquad [\boldsymbol{C}^{(t)}]^{\mathrm{T}} \boldsymbol{X}^{(t)} \leqslant K^{(t)}; \\
\qquad [\boldsymbol{D}^{(t)}]^{\mathrm{T}} \boldsymbol{X}^{(t)} \leqslant W^{(t)}; \\
\qquad [\boldsymbol{E}^{(t)}]^{\mathrm{T}} \boldsymbol{X}^{(t)} \leqslant I^{(t)}; \\
\qquad 0 \leqslant \boldsymbol{X}^{(t)} \leqslant \boldsymbol{F}^{(t)},\ f_i(t) \geqslant 0,\ i = 1,2,\cdots,n;\ t = 1,2,\cdots,T.
\end{cases}
\tag{6-15}
$$

其中$f_i(t)$为第t年i部门的最终产品,它可由(5-3)式、(5-7)式或(5-12)式等确定;$\boldsymbol{B}^{(t)} = (b_1^{(t)}, b_2^{(t)}, \cdots, b_n^{(t)})^{\mathrm{T}}$,$b_i^{(t)}$为第$t$年$i$部门单位产品消耗劳动力的系数;$\boldsymbol{C}^{(t)} = (c_1^{(t)}, c_2^{(t)}, \cdots, c_n^{(t)})^{\mathrm{T}}$,$c_i^{(t)}$为第$t$年$i$部门单位产品消耗某种自然资源的系数;$\boldsymbol{D}^{(t)} = (d_1^{(t)}, d_2^{(t)}, \cdots, d_n^{(t)})^{\mathrm{T}}$,$d_i^{(t)}$为第$t$年$i$部门单位产品消耗某种稀缺资源的系数;$\boldsymbol{E}^{(t)} = (e_1^{(t)}, e_2^{(t)}, \cdots, e_n^{(t)})^{\mathrm{T}}$,$e_i^{(t)}$为第$t$年$i$部门生产单位产品占用资金的系数;$L^{(t)}$,$K^{(t)}$,$W^{(t)}$和$I^{(t)}$分别为第$t$年社会所能提供的劳动力数量、某种自然资源数量、某种稀缺资源的数量和资金的数量.若自然资源或稀缺资源的限制种类不止一种,应分别列出相应的约束条件.

6.3 多目标动态投入产出优化模型①

为制定地区中长期发展规划的需要,那日萨、唐焕文建立了一个多目标动态投入产出优化模型,该模型包含6个目标、4类约束.多目标规划的特点是目标间往往存在一定矛盾.为兼顾各目标,作者利用多目标规划中的目标规划方法来解决该问题;为减少问题中约束条件个数,以各部门产出的增量作为决策变量,使计算的不稳定性得到了较大的改善.

特别地,当只考虑其中一个目标时,相应模型即为一种单目标动态投入产出优化模型.

6.3.1 模型的结构描述

1. 目标函数

目标函数有以下6个:

(1) 规划期内国内生产总值累计最大;

(2) 规划期末年第三产业增加值比重尽可能达到预期目标;

(3) 尽可能达到综合经济平衡;

(4) 规划期内某阶段国内生产总值平均增长速度尽可能达到预期目标;

(5) 能耗最低;

(6) 污染最小.

上述6个目标中,第1个目标是通常考虑最多的目标之一;考虑到第三产业水平是衡量现代社会经济发展程度的重要标志之一,又鉴于我国改革开放前片面推行了工业化政策,忽视非物质生产部门特别是服务业的发展,导致第三产业发展滞后、产业结构失衡的事实,该模型提出了规划期末年第三产业增加值比重尽可能达到预期目标这个目标函数;综合经济平衡是经济系统良性发展的基础,即要求国民经济各部门协调发展,供需平衡,但经济系统实际上是个非均衡系统,因此该模型提出了第3个目标函数;为了使所做的决策符合国家的大政方针,尤其是符合国家经济发展目标,有必要对特定时期规划GDP的增长速度,例如,制定"九五"规划期这一时间段的GDP年平均发展速度,这是第4个目标函数要反映的内容;为了使地区能够可持续发展,必须有效配置资源、节约能源、控制污染、改善环境,该模型提出了最后两个目标函数.

2. 约束条件

约束条件有以下4个:

(1) 动态投入产出平衡约束;

(2) 积累、消费约束;

① 那日萨,唐焕文.一个多目标动态投入产出优化模型及算法(Ⅰ)(Ⅱ).系统工程理论与实践,1998(8,9).

(3) 资源约束；

(4) 变量的非负约束.

其中(1)是基本约束；积累与消费比例关系是经济稳定高效发展的重要关系之一，且总积累、总消费与净流入之和应不超过规划期内的国内生产总值；资源约束包括自然资源、能源及劳动力、资金等方面的限制.

6.3.2　模型的数学描述

1. 目标函数

(1) 规划期内 GDP 累计最大.

$$\max \sum_{t=1}^{T} f(t), \tag{6-16}$$

其中 $f(t)=\sum_{i=1}^{n} X_i(t)-\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(t)X_j(t)$ 为第 t 年的 GDP，T 为规划期.

(2) 规划期末年第三产业增加值尽可能达到预期目标.

$$\min \left| \sum_{i\in N_3} C_i(T)X_i(T)-rf(T) \right|, \tag{6-17}$$

其中 N_3 为第三产业部门指标集，$C_i(T)$为第 T 年末 i 部门的增加值率，即 $C_j(T)=1-\sum_{i=1}^{n} a_{ij}(T)$，$r$ 为规划期末年第三产业增加值占 GDP 的期望比重.

(3) 尽可能达到综合平衡.

$$\min \sum_{t=1}^{T}\sum_{i=1}^{n} (\eta_{-i}(t)+\eta_{+i}(t)), \tag{6-18}$$

其中 $\eta_{-i}(t)$、$\eta_{+i}(t)$分别代表第 t 年 i 部门动态投入产出平衡方程的正、负偏差变量(参见(6-23)式).

(4) 规划期内某时间段内 GDP 期望平均增长速度尽可能达到预期目标.

设这时间段为$[T_1,T_2]$，其中 $T_1\geqslant 1$，$T_2\leqslant T$，该时间段内 GDP 期望平均增长速度为 R，则该目标函数可表述为

$$\min|f_1(T_1-1)(1+R)^{(T_2-T_1+1)}-f(T_2)|, \tag{6-19}$$

(5) 能耗最低.

$$\min \sum_{t=1}^{T} [E^{(t)}]^{\mathrm{T}} X^{(t)}, \tag{6-20}$$

其中 $E^{(t)}=(e_1^{(t)},e_2^{(t)},\cdots,e_n^{(t)})^{\mathrm{T}}$，$e_i^{(t)}$ 为第 t 年 i 部门单位产出的能耗，如水、电、煤等能耗.

(6) 污染最小.

$$\min \sum_{t=1}^{T} [P^{(t)}]^{\mathrm{T}} X^{(t)}, \tag{6-21}$$

其中 $P^{(t)}=(P_1^{(t)},P_2^{(t)},\cdots,P_n^{(t)})^{\mathrm{T}}$，$P_i^{(t)}$ 为第 t 年 i 部门的污染指数，它可通过每万元产出的三废排放量来衡量.

2. 约束条件

(1)动态投入产出平衡约束. 这里采用的是列昂惕夫考虑时变的动态模型,即由(5-7)式经变化得到

$$\boldsymbol{X}^{(0)}=\boldsymbol{A}^{(0)}\boldsymbol{X}^{(0)}+\boldsymbol{B}^{(0)}(\boldsymbol{X}^{(1)}-\boldsymbol{X}^{(0)})+\boldsymbol{Y}^{(0)}+\boldsymbol{\eta}_-(0)-\boldsymbol{\eta}_+(0), \tag{6-22}$$

$$\boldsymbol{X}^{(t)}=\boldsymbol{A}^{(t)}\boldsymbol{X}^{(t)}+\boldsymbol{B}^{(t)}(\boldsymbol{X}^{(t+1)}-\boldsymbol{X}^{(t)})+\boldsymbol{Y}^{(t)}+\boldsymbol{\eta}_-(t)-\boldsymbol{\eta}_+(t),\ t=1,2,\cdots,T, \tag{6-23}$$

其中 $\boldsymbol{\eta}_-(0),\boldsymbol{\eta}_+(0),\boldsymbol{\eta}_-(t),\boldsymbol{\eta}_+(t)$为相应的偏差变向量.

(2) 积累、消费约束.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}^{(t)}(X_j^{(t+1)}-X_j^{(t)})\leqslant f(t)-\sum_{i=1}^{n}Y_i^{(t)},\quad t=1,2,\cdots,T, \tag{6-24}$$

(3) 资源约束.

$$[\boldsymbol{K}_i^{(t)}]^{\mathrm{T}}\boldsymbol{X}^{(t)}\leqslant\overline{\boldsymbol{K}}_i^{(t)},\quad t=1,2,\cdots,T+1;\quad i=1,2,\cdots,m, \tag{6-25}$$

其中 $\boldsymbol{K}_i^{(t)}=(k_{i1}^{(t)},k_{i2}^{(t)},\cdots,k_{in}^{(t)})^{\mathrm{T}}$,$k_{ij}^{(t)}$是第$t$年$j$部门单位产出消耗$i$资源的系数,$\overline{K}_i^{(t)}$是第$t$年$i$资源的最大可提供量,$i=1,2,\cdots,m$.

(4) 非负约束. 设

$$\begin{aligned}&\boldsymbol{X}^{(t+1)}-\boldsymbol{X}^{(t)}\geqslant 0, && t=0,1,\cdots,T,\\ &\boldsymbol{\eta}_+(t)\geqslant 0,\boldsymbol{\eta}_-(t)\geqslant 0, && t=0,1,\cdots,T.\end{aligned} \tag{6-26}$$

6.3.3 一种线性变换下的模型

记

$$\Delta\boldsymbol{X}^{(t)}=\boldsymbol{X}^{(t+1)}-\boldsymbol{X}^{(t)},\ t=1,2,\cdots,T, \tag{6-27}$$

则 $\boldsymbol{X}^{(t+1)}=\boldsymbol{X}^{(t)}+\Delta\boldsymbol{X}^{(t)}=\boldsymbol{X}^{(t-1)}+\Delta\boldsymbol{X}^{(t-1)}+\Delta\boldsymbol{X}^{(t)}=\boldsymbol{X}^{(t-2)}+\Delta\boldsymbol{X}^{(t-2)}+\Delta\boldsymbol{X}^{(t-1)}+\Delta\boldsymbol{X}^{(t)}=\cdots=\boldsymbol{X}^{(0)}+\Delta\boldsymbol{X}^{(0)}+\Delta\boldsymbol{X}^{(1)}+\cdots+\Delta\boldsymbol{X}^{(t)}$,$t=1,2,\cdots,T$,即 $\boldsymbol{X}^{(t)}$可用$\boldsymbol{X}^{(0)},\Delta\boldsymbol{X}^{(0)},\cdots,\Delta\boldsymbol{X}^{(t-1)}$表示,因此将$\Delta\boldsymbol{X}^{(0)},\Delta\boldsymbol{X}^{(1)},\cdots,\Delta\boldsymbol{X}^{(T)}$视作决策变量时,(6-26)式也可改写成$\Delta\boldsymbol{X}^{(t)}\geqslant 0$,$t=0,1,\cdots,T$,再将上述模型中所有目标函数和约束条件表述成$\Delta\boldsymbol{X}^{(0)},\Delta\boldsymbol{X}^{(1)},\cdots,\Delta\boldsymbol{X}^{(T)}$的线性关系式,则问题将减少 $n\times(T+1)$个约束条件,从而减小问题规模. 变换后的目标函数和约束条件如下.

1. 目标函数

(1) $$\max\sum_{t=1}^{T}\sum_{j=1}^{n}C_j(t)[X_j^{(0)}+\Delta X_j^{(0)}+\cdots+\Delta X_j^{(t-1)}]. \tag{6-28}$$

(2) $$\max\sum_{i\in N_3}C_i(T)(X_i^{(0)}+\Delta X_i^{(0)}+\cdots+\Delta X_i^{(T-1)})-r\sum_{j=1}^{n}C_j(T)\cdot[X_j^{(0)}+\Delta X_j^{(0)}+\cdots+\Delta X_j^{(T-1)}]. \tag{6-29}$$

(3) $$\min\sum_{t=1}^{T}\sum_{i=1}^{n}[\eta_{-i}(t)+\eta_{+i}(t)].$$

(4) $$\max\sum_{j=1}^{n}C_j(T_2)[X_j^{(0)}+\Delta X_j^{(0)}+\cdots+\Delta X_j^{(T_2-1)}]-$$

$$\sum_{j=1}^{n} C_j(T_1-1)[X_j^{(0)}+\Delta X_j^{(0)}+\cdots+\Delta X_j^{(T_1-2)}](1+R)^{(T_2-T_1+1)}. \tag{6-30}$$

(5) $$\min \sum_{t=1}^{T}[E^{(t)}]^{\mathrm{T}}(X^{(0)}+\Delta X^{(0)}+\cdots+\Delta X^{(t-1)}). \tag{6-31}$$

(6) $$\min \sum_{t=1}^{T}[P^{(t)}]^{\mathrm{T}}(X^{(0)}+\Delta X^{(0)}+\cdots+\Delta X^{(t-1)}). \tag{6-32}$$

2. 约束条件

(1) $$B^{(0)}\Delta X^{(0)}+\eta_-(0)-\eta_+(0)+Y^{(0)}=(I-A^{(0)})X^{(0)}, \tag{6-33}$$

$$(A^{(t)}-I)(\Delta X^{(0)}+\cdots+\Delta X^{(t-1)}+X^{(0)})+B^{(t)}\Delta X^{(t)}+\eta_-(t)-\eta_+(t)=-Y^{(t)},\quad t=1,2,\cdots,T. \tag{6-34}$$

(2) $$\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^{(t)}\Delta X_j^{(t)} \leqslant \sum_{j=1}^{n} C_j(t)[X_j^{(0)}+\Delta X_j^{(0)}+\cdots+\Delta X_j^{(t-1)}]-\sum_{i=1}^{n} Y_i^{(t)},\quad t=1,2,\cdots,T. \tag{6-35}$$

(3) $$[K_i^{(t)}]^{\mathrm{T}}[X^{(0)}+\Delta X^{(0)}+\cdots+\Delta X^{(t-1)}]\leqslant \overline{K}_i^{(t)},\quad t=1,2,\cdots,T+1;\ i=1,2,\cdots,m. \tag{6-36}$$

(4) $$\Delta X^{(t)}\geqslant 0,\ t=1,2,\cdots,T, \tag{6-37}$$

$$\eta_+(t)\geqslant 0,\ \eta_-(t)\geqslant 0,\quad t=0,1,\cdots,T.$$

上述多目标规划模型的求解,可以通过将各目标加权形成单目标的方法来进行,在模型中直接消耗系数 $A^{(t)}$ 和投资系数 $B^{(t)}$ 须由已有的投入产出表和规划修订得到.

本篇介绍投入产出法的基本内容,同时也尝试介绍一些新内容,限于篇幅和作者的水平,面对日新月异的投入产出技术,总有挂一漏万之感,有许多颇具特色的新方法、新应用没有涉及,作为补偿,这里简单罗列一些,以供读者参考.

比如,由贾芝锡同志研究的地质勘探投入产出模型,由于产出的随机性很大,已有的投入产出技术在这里应用有较大困难,因此,他的研究有特色;由刘起运同志提出的对称模型,其特点是按行计算分配系数,他还研究了该模型与原模型的关系;由陈锡康同志提出的投入占用产出表,表中包含了更多的信息,扩展了分析、预测经济问题的范围;由张守一同志提出的嵌入式投入产出模型及其优化,将投入产出表中一个要考察的部门划分为许多行业,将平衡或优化解嵌入全局投入产出表中,它不仅能反映该部门各行业内部的投入与产出关系,而且能反映该部门各行业与其他部门间的经济联系,能较好地解决部门(行业)编制计划与规划问题;由龚肖宁同志提出的最终需求与收入分配之间的关系,分析了收入结构与消费结构的变化对经济增长、产业结构、平等与效率问题等的巨大影响;由刘树成同志研究的投入产出扩展模型,既反映了产品的运动,又反映了资金的运动,对研究总需求与总供给以及价格等问题有较大的实用价值;由张守一等提出的科技投入产出模型,将固定资产按技术水平、劳动力按文化程度划分为若干等级,用经济计量方法估计广

义生产函数的参数,将它们正规化,按贡献率分解为劳动报酬、剩余产品或净产品,从而测算科技进步对经济增长的贡献;由贺铿同志负责研究的湖南省岳阳县信息投入产出表及1987年全国信息投入产出表,由列昂惕夫研究的全球环境保护模型等等,都是新的成果.另外,尚有将投入产出技术与经济计量学相结合、与系统动力学相结合、与马尔可夫链及与动态控制模型相结合等有关投入产出分析研究与应用的成果.

投入产出技术仍然有许多问题值得研究.比如,一般均衡论是投入产出法的理论基础之一,但国民经济在运作过程中通常呈非均衡状态.除了部分优化模型外,大部分模型没有考虑非均衡这一状况;经济系统的运作本质是非线性的,但这方面研究还较少.此外,投入产出技术与灰色系统的结合、与模糊数学的结合、与对策论的结合,以及直接消耗系数阵、投资系数的修订等问题,均有待进一步探讨.

参 考 文 献

1 钟契夫主编.投入产出分析(修订本).北京:中国财政经济出版社,1993.
2 联合国统计局编.投入产出表和分析.萧嘉魁,周逸江译校.北京:中国社会科学出版社,1981.
3 李秉全.投入产出技术与企业管理现代化.北京:科学出版社,1988.
4 张守一,葛新权.中国宏观经济理论·模型·预测.北京:社会科学文献出版社,1995.
5 陈锡康.投入产出方法.北京:人民出版社,1983.
6 薛俊杰,李春森.投入产出法教程.大连:东北财经大学出版社,1992.
7 (美)威廉 H 密尔涅克著.投入-产出分析基础理论.秋同译.北京:中国社会科学出版社,1980.
8 (英)欧考纳 R,亨利 E W 著.投入产出分析及其应用.夏绍玮,赵纯均译.北京:清华大学出版社,1984.
9 (前苏联)瓦西里·列昂惕夫著.投入产出经济学.崔书香译.北京:商务印书馆,1980.
10 庞皓,向蓉美.投入产出分析.成都:西南财经大学出版社,1989.
11 赵新良等.动态投入产出.沈阳:辽宁人民出版社,1988.

·经济数学卷·

第12篇

线性控制系统理论

编　者　陈彭年　秦化淑
审校者　黄　琳

目　录

引　　言

线性控制系统理论主要研究线性控制系统的结构和控制性质，以及依据这些性质给出控制系统的设计原理和方法.它在理论和应用两方面都很重要.一方面，它是控制理论其他分支的理论基础；另一方面，它被广泛地应用于自动控制工程的设计，并成为其他领域控制问题分析和综合的有用工具.

在线性控制系统理论的发展史上，通常称20世纪60年代以前的为经典线性系统理论.经典线性系统理论以传递函数作为系统的数学模型，以常微分方程和频率响应法为工具，主要研究单输入单输出系统的输入输出特性.1960年前后，卡尔曼(R.E.Kalman)提出了线性控制系统的状态空间方法，以及能控和能观测两个重要概念，奠定了现代线性控制系统理论的基础.现代线性系统理论以状态方程和输出方程作为系统的数学模型，以线性代数和常微分方程等为主要研究工具，研究内容不仅涉及多输入多输出系统的输入输出特性，还重点研究多变量系统的内部性质，使线性控制系统的分析综合建立在严格的数学理论基础之上.

本篇主要介绍连续时间线性控制系统状态空间方法.

1　线性控制系统的数学描述

在控制系统的分析和设计中，第一步是建立数学模型，以对其进行适当的数学描述.对于线性控制系统，有多种数学描述方法，但最常见的是传递函数法和状态空间法.

1.1　传递函数描述方法

1.1.1　传递函数概念

传递函数是用来表示系统输入输出关系的一种数学表达式.

定义1　设如图1-1所示，Σ是受控制系统，u是系统的输入，y是系统的输出，Σ是单输入单输出系统，则在系统初始条件为零的条件下，系统输出y的拉普拉斯(P.S.Laplace)变换$Y(s)$与其输入u的拉普拉斯变换$U(s)$之比

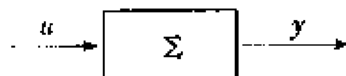


图 1-1

$$G(s)=\frac{Y(s)}{U(s)}\tag{1-1}$$

称为系统Σ的**传递函数**.

$$Y(s) = G(s)U(s) \tag{1-2}$$

称为系统 Σ 的传递函数描述.

定义 2　当系统 Σ 有 $m(m \geqslant 1)$ 个输入和 $p(p \geqslant 1)$ 个输出时,任一个输入和任一个输出之间都有一个传递函数,即共有 mp 个传递函数.若将这 mp 个传递函数按一定顺序排列成一个 $p \times m$ 矩阵 $\boldsymbol{G}(s)$,则称 $\boldsymbol{G}(s)$ 为系统的**传递函数矩阵**.

1.1.2　单输入单输出定常系统的传递函数

传递函数主要用于定常线性系统的描述.设一个单输入单输出定常线性系统的输入输出关系由下列 n 阶常微分方程表示:

$$\begin{aligned} & y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1 y'(t) + a_0 y(t) \\ & = b_m u^{(m)}(t) + b_{m-1}u^{(m-1)}(t) + \cdots + b_1 u'(t) + b_0 u(t). \end{aligned} \tag{1-3}$$

其中 $y(t)$ 是系统的输出;$u(t)$ 是系统的输入;t 表示时间;$y^{(k)}(t)$, $u^{(l)}(t)$ 分别是 $y(t)$, $u(t)$ 对 t 的 k, l 次导数;a_i, b_j 都是常数,$i = 0,1,2,\cdots,n-1$, $j = 0,1,2,\cdots,m$.

不失一般性,设初始时刻 $t_0 = 0$,系统(1-3)式的初始条件为零,即 $y(0) = y'(0) \cdots = y^{(n-1)}(0) = 0$, $u(0) = u'(0) = \cdots = u^{(m-1)}(0) = 0$.在零初始条件下,对方程(1-3)式作拉普拉斯变换后,得

$$\begin{aligned} & (s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0)Y(s) \\ & = (b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0)U(s), \end{aligned}$$

其中 $Y(s)$ 和 $U(s)$ 分别是 $y(t)$ 和 $u(t)$ 的拉普拉斯变换(式),s 为拉普拉斯算符.

于是系统(1-3)式的传递函数为

$$G(s) = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0}. \tag{1-4}$$

如果 $m \leqslant n$,则 $G(s)$ 是真有理分式.此时称系统(1-3)式是物理能实现的.在线性系统理论里,只研究物理能实现的系统.

多输入多输出定常线性系统传递函数矩阵表达式见下一节状态空间描述法.

1.1.3　系统的特征多项式、极点和零点

定义 3　(1-4)式中分母多项式

$$p(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 \tag{1-5}$$

称为系统(1-3)式的**特征多项式**.

定义 4　方程

$$s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0 \tag{1-6}$$

称为系统(1-3)式的特征方程.

定义 5　特征多项式的零点,或者特征方程的根,称为系统的**极点**.

定义 6　方程

$$b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0 = 0 \tag{1-7}$$

的根称为系统(1-3)式的**零点**.

如果系统(1-3)式有相同的零点和极点,则称其有**零极相消**.

定义 7 零极相消后剩下的极点和零点分别称为传递函数 $G(s)$的极点和零点.

定义 8 如果系统(1-3)式的所有零点和极点都在开的左半复平面上,则称其为最小相位系统.

1.2 状态空间描述方法

1.2.1 状态变量和状态空间

状态空间描述法是建立在状态变量概念之上的空间描述方法.系统在 t_0 时刻的状态是指系统在 t_0 时刻的信息量,它与从 t_0 时刻起作用于系统中的输入量一起,唯一地确定系统在 $t \geqslant t_0$ 时的动力学行为.

定义 9 称一组变量为一个系统的状态变量,若该组变量是描述该系统的动力学行为所需的一组最少的独立变量.由系统的状态变量组成的向量称为系统的状态向量.状态向量取值的有限维欧几里德(Euclidean)空间称为系统的状态空间.

1.2.2 系统的状态空间描述法

用状态空间方法描述的线性系统为

$$\dot{x} = A(t)x + B(t)u, \tag{1-8}$$

$$y = C(t)x + D(t)u. \tag{1-9}$$

其中 $x \in \mathbf{R}^n$,叫做系统的状态向量;$u \in \mathbf{R}^m$,叫做系统的控制(或输入)向量;$y \in \mathbf{R}^p$,叫做系统的量测(或输出)向量;$A(t) \in \mathbf{R}^{n \times n}$,叫做状态矩阵;$B(t) \in \mathbf{R}^{n \times m}$,叫做控制(或输入)矩阵;$C(t) \in \mathbf{R}^{p \times n}$,叫做量测(或输出)矩阵;$D(t) \in \mathbf{R}^{p \times m}$,叫做前馈矩阵;$t$ 是时间变量,这些矩阵中的元都是 t 的分段连续函数.

定义 10 方程(1-8)式称为系统的状态方程,方程(1-9)式称为系统的量测方程(或输出方程).

定义 11 状态向量 x 的维数,或者状态变量的个数,称为系统的阶.一个 n 阶系统对应的状态空间是 n 维的.

定义 12 如果在系统(1-8)式和(1-9)式中,矩阵 $A(t)$,$B(t)$,$C(t)$和 $D(t)$都是常值矩阵,则称该系统是定常的或时不变的,否则称为时变的.如果 $m = 1$,则称其为单输入的;如果 $p = 1$,则称其为单输出的;如果 $m = p = 1$,则称其为单输入单输出系统.

定义 13 方程

$$\dot{x} = A(t)x \tag{1-10}$$

称为状态方程(1-8)式的自由系统.

1.2.3 定常线性系统

1. 系统描述

用状态空间描述的定常线性系统为

$$\dot{x} = Ax + Bu, \tag{1-11}$$
$$y = Cx + Du. \tag{1-12}$$

其中 $x \in \mathbf{R}^n$；$u \in \mathbf{R}^m$；$y \in \mathbf{R}^p$；$A \in \mathbf{R}^{n\times n}$，$B \in \mathbf{R}^{n\times m}$，$C \in \mathbf{R}^{p\times n}$和$D \in \mathbf{R}^{p\times m}$皆为常值矩阵.

2. 传递函数矩阵

定义 14　称 $p\times m$ 有理分式矩阵

$$G(s) = C(sI_n - A)^{-1}B + D \tag{1-13}$$

为系统(1-11)式和(1-12)式的传递函数矩阵,其中 I_n 为 n 阶单位矩阵.

3. 系统的特征多项式和极点

定义 15　多项式

$$p(s) = \det(sI_n - A) \tag{1-14}$$

称为系统(1-11)式和(1-12)式的特征多项式.

定义 16　代数方程

$$\det(sI_n - A) = 0 \tag{1-15}$$

称为系统(1-11)式和(1-12)式的特征方程.

特征多项式的零点称为系统的极点.

4. 系统零点

定义 17　设

$$\mathbf{C}_I = \{s \mid \operatorname{rank}[sI_n - A \;\; B] < n, s \in \mathbf{C}\}, \tag{1-16}$$

$$\mathbf{C}_0 = \left\{s \mid \operatorname{rank}\begin{bmatrix} sI_n - A \\ C \end{bmatrix} < n, s \in \mathbf{C}\right\}, \tag{1-17}$$

$$\mathbf{C}_\Sigma = \left\{s \mid \operatorname{rank}\begin{bmatrix} sI_n - A & -B \\ -C & -D \end{bmatrix} < n + \min\{m, p\}, s \in \mathbf{C}\right\}, \tag{1-18}$$

其中 C 表示复平面,则复数集合 $\mathbf{C}_I$,$\mathbf{C}_0$ 和 $\mathbf{C}_\Sigma$ 分别称为系统(1-11)式和(1-12)式的输入解耦零点、输出解耦零点和零点.

系统传输零点的概念见参考文献[1].

1.3　两种描述方法的比较

本节对系统的传递函数描述法和状态空间描述法的优点和不足之处做一简单对比.

首先,传递函数描述法仅仅描述了系统的输入输出关系,反映了系统的外部特性,但却没有反映系统的内部结构.特别是当系统出现零极相消时,传递函数描述法就无法反映系统的动力学特性了.然而,用状态空间法描述系统,不仅能反映系统的输入输出关系,也刻画了系统内部结构的动力学行为.因此,它更加完善地描述了系统.

其次,由于两种方法使用的数学工具不同,因此,适用的范围也就有所不同.传递函数描述法一般仅适用于定常线性系统,而状态空间描述法既可以描述定常线

性系统，又可描述时变线性系统. 其实状态空间描述法也适用于描述各种非线性系统.

再次，对于很复杂的系统，建立它的状态方程和量测方程是困难和麻烦的，这时借助于对系统的输入、输出信号的测量，采用系统辨识方法可能会比较容易地确定系统的传递函数或脉冲响应函数，而要辨识系统完整的状态方程和量测方程有时是难于做到的.

最后，传递函数描述法是经典控制理论中进行分析和综合的基础，它可借助于一些简单的方法完成反馈控制系统的设计；状态空间描述法是现代控制理论用于系统设计的基础，它能解决那些经典理论所不能处理的问题. 不过，用状态空间描述法设计系统时，往往要借助于计算机才能完成其设计工作.

总之，传递函数法和状态空间法各有所长，目前，它们在线性系统的分析和综合中都起着重要作用.

1.4 线性系统的等价性

1.4.1 代数等价

定义 18 设有定常线性系统(1-11)式和(1-12)式，作坐标变换

$$x = P\bar{x}, \quad \det P \neq 0,$$

则在新坐标系下，系统(1-11)式和(1-12)式变为

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u, \tag{1-19}$$

$$y = \bar{C}\bar{x} + \bar{D}u, \tag{1-20}$$

其中

$$\bar{A} = P^{-1}AP, \bar{B} = P^{-1}B,$$
$$\bar{C} = CP, \bar{D} = D.$$

称系统(1-19)式和(1-20)式为与系统(1-11)式和(1-12)式的**代数等价系统**.

显然，系统(1-11)式和(1-12)式也是系统(1-19)式和(1-20)式的代数等价系统.

时变线性系统的代数等价性可作类似的定义，可参见文献[2].

1.4.2 代数等价系统的性质

代数等价系统主要性质为

1° 有相同的特征多项式，从而有相同的极点；

2° 有相同的系统输入解耦零点、输出解耦零点和系统零点；

3° 有相同的传递函数矩阵.

从上面 1°，2°和 3°这些性质看，代数等价系统本质上是相同的.

2 线性控制系统的能控性和能观测性

2.1 能 控 性

2.1.1 定义

考虑线性控制系统

$$\begin{cases} \dot{x} = A(t)x + B(t)u, \\ y = C(t)x, \end{cases} \tag{2-1}$$

其中各符号的含义与系统(1-8)式和(1-9)式的相同,只不过这里取 $D(t)=0$.

用 $x(t;t_0,x_0,u)$表示系统(2-1)式在 t_0 时过点 x_0 且在控制 $u=u(t)$的作用下的解,即

$$x(t;t_0,x_0,u) = \Phi(t,t_0)x_0 + \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)\mathrm{d}\tau,$$

其中 $\Phi(t,t_0)$是自由系统

$$\dot{x} = A(t)x \tag{2-2}$$

的状态转移矩阵.

定义 1 如果在时刻 t_0,对于任意的 $x_0 \in \mathbf{R}^n$,总能找到时刻 $t_1 > t_0$ 和定义在$[t_0,t_1]$上的容许控制 $u=u(t)$,使得

$$x(t_1;t_0,x_0,u(\cdot)) = 0,$$

那么称系统(2-1)式在 t_0 时完全能控,简称**能控**或称系统具有**能控性**.如果系统(2-1)式在区间$[t_0,T]$的每个时刻都完全能控,则称它在$[t_0,T]$上完全能控.

在线性系统理论中,容许控制一般取分段连续函数.

从定义 1 可知,时刻 t_1 同给定的初始状态 x_0 有关,但由于状态空间是有限维的,如果系统(2-1)式可控,那么一定存在一个不依赖于初始状态 x_0 的 t_1.

定义 2 如果在时刻 t_0,对于任意的 $x_1 \in \mathbf{R}^n$,总能找到时刻 $t_1 > t_0$ 和定义在$[t_0,t_1]$上的容许控制 $u=u(t)$,使得 $x(t_1;t_0,0,u)=x_1$,那么称系统(2-1)式在时刻 t_0 是完全能达的,简称**能达**.

在能达性的定义 2 中,系统在 t_0 时的状态是零.

定理 1 对于线性系统,能控性和能达性是等价的,即系统(2-1)式在 t_0 时能控的充分必要条件是其在 t_0 时能达.

2.1.2 时变线性系统能控性的判据

定理 2 系统(2-1)式在 t_0 时能控的充分必要条件是存在 $t_1 > t_0$,使得矩阵

$$W(t_1,t_0) = \int_{t_0}^{t_1} \Phi(t_1,\tau)B(\tau)B^{\mathrm{T}}(t)\Phi^{\mathrm{T}}(t_1,\tau)\mathrm{d}\tau$$

非奇异,其中 $\boldsymbol{\Phi}(t,\tau)$为系统(2-2)式的状态转移矩阵.

定理 3　假设在系统(2-1)式中,矩阵 $\boldsymbol{A}(t)$和 $\boldsymbol{B}(t)$的每个元在$[t_0,+\infty)$上分别是 $n-2$ 阶和 $n-1$ 阶连续可微.记

$$\boldsymbol{B}_1(t)=\boldsymbol{B}(t),$$

$$\boldsymbol{B}_i(t)=-\boldsymbol{A}(t)\boldsymbol{B}_{i-1}(t)+\dot{\boldsymbol{B}}_{i-1}(t)\quad(i=2,3,\cdots,n).$$

设

$$\boldsymbol{Q}(t)=[\boldsymbol{B}_1(t),\boldsymbol{B}_2(t),\cdots,\boldsymbol{B}_n(t)],$$

如果存在 $t_1>t_0$,使得 $\operatorname{rank}\boldsymbol{Q}(t_1)=n$,那么系统(2-1)式在 t_0 时能控.

例 1　判别时变系统

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix}=\begin{bmatrix}0 & t\\ 0 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}+\begin{bmatrix}0\\ 1\end{bmatrix}u(t),\tag{2-3}$$

在 $t=0$ 时的能控性.

解　方法 1（由定理 2 判别）　经计算可知,系统(2-3)式的自由系统的状态转移矩阵

$$\boldsymbol{\Phi}(t,\tau)=\begin{bmatrix}1 & \frac{1}{2}(t-\tau)^2\\ 0 & 1\end{bmatrix}.$$

能控性矩阵

$$\begin{aligned}\boldsymbol{W}(t_1,0)&=\int_0^{t_1}\boldsymbol{\Phi}(t_1,\tau)\begin{bmatrix}0\\ 1\end{bmatrix}[0,1]\boldsymbol{\Phi}^T(t_1,\tau)\mathrm{d}\tau\\ &=\int_0^{t_1}\begin{bmatrix}1 & \frac{1}{2}(t_1-\tau)^2\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix}[0,1]\begin{bmatrix}1 & 0\\ \frac{1}{2}(t_1-\tau)^2 & 1\end{bmatrix}\mathrm{d}\tau\\ &=\int_0^{t_1}\begin{bmatrix}\frac{1}{4}(t_1-\tau)^4 & \frac{1}{2}(t_1-\tau)^2\\ \frac{1}{2}(t_1-\tau)^2 & 1\end{bmatrix}\mathrm{d}\tau\\ &=\begin{bmatrix}\frac{1}{20}t_1^5 & \frac{1}{6}t_1^3\\ \frac{1}{6}t_1^3 & t_1\end{bmatrix}.\end{aligned}$$

$$\det\boldsymbol{W}(t_1,0)=\frac{1}{45}t_1^6>0\qquad(t_1>0).$$

因此,依定理 2,系统(2-3)式在 $t_0=0$ 时能控.

方法 2（由定理 3 判别）　设

$$\boldsymbol{B}_1(t)=\begin{bmatrix}0\\ 1\end{bmatrix},$$

$$\boldsymbol{B}_2(t)=-\boldsymbol{A}(t)\boldsymbol{B}_1(t)+\dot{\boldsymbol{B}}_1(t)=\begin{bmatrix}-t\\ 0\end{bmatrix}.$$

于是

$$\operatorname{rank}[\boldsymbol{B}_1(t),\quad \boldsymbol{B}_2(t)]=\operatorname{rank}\begin{bmatrix}0 & -t\\ 1 & 0\end{bmatrix}=2\quad(t>0).$$

因此，依定理3，系统(2-3)式在 $t_0=0$ 时能控.

2.1.3　定常线性系统的能控性

能控性与输出方程无关，设定常线性系统为

$$\dot{x}=Ax+Bu,\tag{2-4}$$

其中 $x\in\mathbf{R}^n$，$u\in\mathbf{R}^m$；$A\in\mathbf{R}^{n\times n}$，$B\in\mathbf{R}^{n\times m}$.

容易看到，定常线性系统的能控性与初始时刻 t_0 无关，因此，在讨论能控性时，总将其初始时刻 t_0 省去.

定理4（秩判据）　系统(2-4)式能控的充分必要条件为

$$\operatorname{rank}[B,AB,\cdots,A^{n-1}B]=n.$$

记

$$Q_c=[B,AB,\cdots,A^{n-1}B].\tag{2-5}$$

Q_c 称为系统(2-4)式的能控性矩阵.

定理4表明，要判别系统(2-4)式是否能控，只须判其能控性矩阵的秩是否为 n 即可. 这为能控性提供了一个实际可行的判别法.

推论1　设系统(2-4)式是单输入的，即 $B=b\in\mathbf{R}^n$，则系统(2-4)式能控的充分必要条件为

$$\det[b,Ab,\cdots,A^{n-1}b]\neq 0.\tag{2-6}$$

例2　惯性导航系统的罗经回路方程为

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & g & 0\\ -1/R & 0 & -\Omega\cos\varphi\\ 0 & \Omega\cos\varphi & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}u,\tag{2-7}$$

$$y=x_1,$$

其中 R，g，Ω 和 φ 分别表示地球平均半径、重力加速度、地球自转角速度和地理纬度. 令 φ 是常数. 试问此系统能否可控？

解　经计算，系统(2-7)式的能控性矩阵为

$$Q_c=\begin{bmatrix}0 & 0 & -g\Omega\cos\varphi\\ 0 & -\Omega\cos\varphi & 0\\ 1 & 0 & -\Omega^2\cos^2\varphi\end{bmatrix}.$$

不难验证，只要 $\varphi\neq\pm\frac{1}{2}\pi$，就有 $\det Q_c=-g\Omega^2\cos^2\varphi\neq 0$. 根据推论1(或定理4)，系统(2-7)式是能控的. 这就是说，系统不处于南、北两极时，一定是能控的.

推论2　设 A 的最小多项式的次数为 k，则系统(2-4)式能控的充分必要条件是

$$\operatorname{rank}[B,AB,\cdots,A^{k-1}B]=n.\tag{2-8}$$

推论3　设 $\operatorname{rank}B=m$，则系统(2-4)式能控的充分必要条件是

$$\operatorname{rank}[B,AB,\cdots,A^{n-m}B]=n.\tag{2-9}$$

定理5（PBH 判据①） 系统(2-4)式能控的充分必要条件是

$$\operatorname{rank}[s\boldsymbol{I}_n-\boldsymbol{A},\boldsymbol{B}]=n \quad (s\in\sigma(\boldsymbol{A})),\tag{2-10}$$

其中 $\sigma(\boldsymbol{A})$是矩阵 $\boldsymbol{A}$ 的谱集.

推论 4 系统(2-4)式能控的充分必要条件为

$$\operatorname{rank}[s\boldsymbol{I}_n-\boldsymbol{A},\boldsymbol{B}]=n \quad (s\in\mathcal{C}),\tag{2-11}$$

其中 $\mathcal{C}$ 表示复平面.

由于对于任意的 $s\in\mathcal{C}$，必有

$$\operatorname{rank}[s\boldsymbol{I}_n-\boldsymbol{A},\boldsymbol{B}]\leqslant n,$$

推论 4 实际上是说，系统(2-4)式能控的充分必要条件是它没有输入解耦零点.

下面介绍能控性的几何判据. 为此，设 $\mathscr{B}=\operatorname{Im}\boldsymbol{B}$，以及

$$\langle \boldsymbol{A}\mid\mathscr{B}\rangle=\mathscr{B}+\boldsymbol{A}\mathscr{B}+\cdots+\boldsymbol{A}^{n-1}\mathscr{B}.\tag{2-12}$$

能够证明：$\langle \boldsymbol{A}\mid\mathscr{B}\rangle$是系统(2-4)式的能控子空间（参见文献[1]）.

定理 6（几何判据） 系统(2-4)式能控的充分必要条件为

$$\langle \boldsymbol{A}\mid\mathscr{B}\rangle=\mathbf{R}^n.\tag{2-13}$$

例 3 试问定常线性系统

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}=\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}1\\ 1\end{bmatrix}u\tag{2-14}$$

是否能控？如果不能控，试求其能控子空间.

解 本题中，$n=2$，并且

$$\boldsymbol{B}=\begin{bmatrix}1\\ 1\end{bmatrix},\quad \mathscr{B}=\operatorname{Im}\begin{bmatrix}1\\ 1\end{bmatrix}=\operatorname{span}\left\{\begin{bmatrix}1\\ 1\end{bmatrix}\right\},$$

$$\boldsymbol{A}\mathscr{B}=\operatorname{span}\{\boldsymbol{AB}\}=\operatorname{span}\left\{\begin{bmatrix}-1\\ -1\end{bmatrix}\right\}.$$

于是有

$$\begin{aligned}\langle \boldsymbol{A}\mid\mathscr{B}\rangle&=\mathscr{B}+\boldsymbol{A}\mathscr{B}\\ &=\operatorname{span}\left\{\begin{bmatrix}1\\ 1\end{bmatrix}\right\}+\operatorname{span}\left\{\begin{bmatrix}-1\\ -1\end{bmatrix}\right\}\\ &=\operatorname{span}\left\{\begin{bmatrix}1\\ 1\end{bmatrix}\right\}\\ &\neq\mathbf{R}^2.\end{aligned}$$

根据定理 6，系统(2-14)式不能控. 其能控子空间为

$$\langle \boldsymbol{A}\mid\mathscr{B}\rangle=\operatorname{span}\left\{\begin{bmatrix}1\\ 1\end{bmatrix}\right\}.$$

---

① 以波波夫(Popov)、贝尔维奇(Belevitch)和豪塔斯(Hautus)命名.

2.2 能观测性

2.2.1 能观测性定义

定义3 假设系统(2-1)式中,$t_0 \in \mathbf{R}$.如果存在时刻 $t_1 > t_0$,使得通过量测在时间间隔$[t_0, t_1]$上的系统输出 $\boldsymbol{y}(t)$和已知的控制输入 $\boldsymbol{u}(t)$,能够唯一地决定出在初始时刻 t_0 的初始状态 $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$,那么称系统(2-1)式在 t_0 时刻是完全能观测的,或称系统具有能观测性.

定义4 设 $\boldsymbol{x}_0$ 是系统(2-1)式在初始时刻 t_0 时的状态.如果当 $t \geqslant t_0$ 时,有 $\boldsymbol{u}(t) \equiv 0$,这时系统的输出恒为零,即 $\boldsymbol{y}(t) \equiv 0$,那么称这个状态 $\boldsymbol{x}_0$ 在时刻 t_0 是不能观测的.

容易看到,t_0 时刻不能观测状态全体形成 $\mathbf{R}^n$ 中一个线性子空间.此子空间称为系统(2-1)式在时刻 t_0 的不能观测子空间.

可以证明:系统(2-1)式在时刻 t_0 能观测的充分必要条件是其在 t_0 时刻没有不能观测的状态.

2.2.2 时变线性系统能观测性判据

定理7 系统(2-1)式在时刻 t_0 能观测的充分必要条件是,存在 $t_1 > t_0$,使得矩阵

$$\boldsymbol{M}(t_1, t_0) = \int_{t_0}^{t_1} \boldsymbol{\Phi}^{\mathrm{T}}(\tau, t_0)\boldsymbol{C}^{\mathrm{T}}(\tau)\boldsymbol{C}(\tau)\boldsymbol{\Phi}(\tau, t_0)\mathrm{d}\tau \qquad (2\text{-}15)$$

非奇异,其中 $\boldsymbol{\Phi}(\tau, t)$是自由系统(2-2)式的状态转移矩阵.

定理8 假设系统(2-1)式中,$t_0 \in \mathbf{R}$;并假定 $\boldsymbol{A}(t)$和 $\boldsymbol{C}(t)$的元素在$[t_0, +\infty)$上分别是 $n-2$ 阶和 $n-1$ 阶连续可微.记

$$\boldsymbol{C}_1(t) = \boldsymbol{C}(t),$$

$$\boldsymbol{C}_i(t) = \boldsymbol{C}_{i-1}(t)\boldsymbol{A}(t) + \dot{\boldsymbol{C}}_{i-1}(t) \quad (i = 2, 3, \cdots, n).$$

设

$$\boldsymbol{R}(t) = \begin{bmatrix} \boldsymbol{C}_1(t) \\ \boldsymbol{C}_2(t) \\ \vdots \\ \boldsymbol{C}_n(t) \end{bmatrix}.$$

如果存在 $t_1 > t_0$,使得

$$\mathrm{rank}\boldsymbol{R}(t_1) = n,$$

则系统(2-1)式在 t_0 时能观测.

2.2.3 能控性和能观测性的对偶原理

从系统的能控性矩阵 $\boldsymbol{W}(t_1, t_0)$和能观性矩阵 $\boldsymbol{M}(t_1, t_0)$的结构看,它们在形式

上有某种类似之处．事实上，这种结构的相似性是由它们的对偶原理决定的．对偶原理是由卡尔曼提出来的．为讨论对偶原理，先引入系统(2-1)式的对偶系统．

定义 5 称线性系统

$$\dot{\psi} = -A^{T}(t)\psi + C^{T}(t)v, \qquad (2\text{-}16)$$
$$z = B^{T}(t)\psi,$$

为系统(2-1)式的对偶系统．其中矩阵 $A(t)$, $B(t)$和 $C(t)$分别为系统(2-1)式的状态矩阵、控制矩阵和量测矩阵；ψ 是对偶系统的状态向量，ψ^{T} 也称为系统(2-1)式状态向量的协态向量；v 和 z 分别是对偶系统的输入向量和量测向量．

定理 9（对偶原理） 系统(2-1)式在时刻 t_0 能控的充分必要条件是其对偶系统(2-16)式在时刻 t_0 能观测．系统(2-1)式在时刻 t_0 能观测的充分必要条件是其对偶系统(2-16)式在时刻 t_0 能控．

2.2.4 定常线性系统能观测性判据

线性系统的能观测性同系统的输入无关．设定常线性系统为

$$\dot{x} = Ax, \quad y = Cx. \qquad (2\text{-}17)$$

其中 $x \in \mathbf{R}^{n}$, $y \in \mathbf{R}^{p}$; $A \in \mathbf{R}^{n \times n}$, $C \in \mathbf{R}^{p}$．

和定常线性系统的能控性一样，定常线性系统的能观测性同初始时刻 t_0 无关．因此在讨论能观性时，也省去"在时刻 t_0"的说法．

下面的能观性判据可以直接证明，也可以根据对偶原理，由 2.2.3 小节中的能控性判据导出．

定理 10 系统(2-17)式能观测的充分必要条件为

$$\operatorname{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n. \qquad (2\text{-}18)$$

推论 5 设系统(2-17)式为单输出，即 $C = c \in \mathbf{R}^{1 \times n}$，则系统(2-17)式能观测的充分必要条件为

$$\det\begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} \neq 0. \qquad (2\text{-}19)$$

例 4 研究系统(2-7)式的能观测性．

解 在系统(2-7)式中，$n = 3$，并且

$$A = \begin{bmatrix} 0 & g & 0 \\ -1/R & 0 & -\Omega\cos\varphi \\ 0 & \Omega\cos\varphi & 0 \end{bmatrix},$$
$$c = [1, 0, 0].$$

因此，可得

$$cA=[0,g,0],$$
$$cA^2=[-g/R,0-g\Omega\cos\varphi].$$

计算行列式

$$\det\begin{bmatrix}c\\cA\\cA^2\end{bmatrix}=\det\begin{bmatrix}1&0&0\\0&g&0\\-g/R&0&-g\Omega\cos\varphi\end{bmatrix}=-g^2\Omega\cos\varphi.$$

如果 $\varphi\neq\frac{1}{2}\pi$,则该行列式不为零.由推论 5(或定理 10)可知,系统(2-7)式能观测.

推论 6　设 A 的最小多项式的次数为 k,则系统(2-17)式能观测的充分必要条件为

$$\text{rank}\begin{bmatrix}C\\CA\\\vdots\\CA^{k-1}\end{bmatrix}=n. \tag{2-20}$$

推论 7　设 $\text{rank}\,C=p$,则系统(2-17)式能观测的充分必要条件为

$$\text{rank}\begin{bmatrix}C\\CA\\\vdots\\CA^{n-p}\end{bmatrix}=n. \tag{2-21}$$

定理 11　系统(2-17)式能观测的充分必要条件为

$$\text{rank}\begin{bmatrix}sI_n-A\\C\end{bmatrix}=n\quad(s\in\sigma(A)), \tag{2-22}$$

其中 I_n 为 n 阶单位矩阵,$\sigma(A)$为 A 的谱集.

推论 8　系统(2-17)式能观测的充分必要条件为

$$\text{rank}\begin{bmatrix}sI_n-A\\C\end{bmatrix}=n\quad(s\in C), \tag{2-23}$$

其中 C 表示复平面.

定理 12　系统(2-17)式能观测的充分必要条件为

$$\bigcap_{i=0}^{n-1}\ker(CA^i)=\varnothing, \tag{2-24}$$

其中 $\ker(CA^i)$表示矩阵 CA^i 的化零子空间;$\varnothing$表示空集.

例 5　判别系统

$$\dot{x}=Ax,\quad y=Cx \tag{2-25}$$

的能观测性,其中 $x\in\mathbf{R}^4$,$y\in\mathbf{R}^2$,且 A,C 分别为

$$A=\begin{bmatrix}0&1&0&0\\3\omega^2&0&0&2\omega\\0&0&0&1\\0&-2\omega&0&0\end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

其中 ω 为常数.

解 经计算

$$\ker C = \operatorname{span}\left\{\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right\},$$

$$\ker CA = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right\},$$

因此,有

$$\bigcap_{i=0}^{3} \ker(CA^i) \subset \ker C \cap \ker CA$$

$$= \operatorname{span}\left\{\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right\} \cap \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right\} = \varnothing.$$

依定理 12,对于所有 $\omega \in \mathbf{R}$,系统(2-25)式能观测.

2.3 定常线性系统的能稳性和能检测性

2.3.1 能稳性

定义 6 称线性系统((2-4)式)

$$\dot{x} = Ax + Bu$$

是能稳的(具有**能稳性**),如果存在状态反馈

$$u = Kx, \tag{2-26}$$

其中 $K \in \mathbf{R}^{m \times n}$ 使得闭环系统

$$\dot{x} = (A + BK)x \tag{2-27}$$

渐近稳定,即 $A + BK$ 的特征值都有负实部.

当系统(2-4)式能稳时,称有序矩阵对(A, B)能稳.

定理 13 (A, B)能稳的充分必要条件为

$$\operatorname{rank}[sI_n - A, B] = n \qquad (s \in \sigma(A) \cap C^+), \tag{2-28}$$

其中 I_n 为 n 阶单位矩阵,$\sigma(A)$为 A 的谱集,C^+ 为闭的右半复平面.

推论 9 (A, B)能稳的充分必要条件为

$$\operatorname{rank}[sI_n - A, B] = n \qquad (s \in C^+). \tag{2-29}$$

设 A 的最小多项式为 α，将 α 因式分解为

$$\alpha=\alpha^{-}\alpha^{+},$$

其中 $\alpha^{-}(\alpha^{+})$ 的复零点属于 $C^{-}(C^{+})$，C^{-} 为开的左半复平面.子空间 $\ker\alpha^{+}(A)$ 叫做 A 的"不稳定振型"子空间.

定理 14　(A,B) 能稳的充分必要条件为

$$\ker\alpha^{+}(A)\subset\langle A\mid\mathscr{B}\rangle. \tag{2-30}$$

此定理称为能稳性的几何判据.

例 6　设所论系统的系统矩阵和控制矩阵分别为

$$A=\begin{bmatrix}0&1&0&0&0\\0&0&0&0&0\\0&0&-2&0&0\\0&0&0&0&1\\0&0&0&0&0\end{bmatrix},\quad B=\begin{bmatrix}0&0\\1&1\\0&0\\0&0\\1&-1\end{bmatrix}.$$

判别 (A,B) 是否能稳.

解　容易求得 A 的最小多项式为

$$\alpha(s)=s^2(s+2).$$

由此，得

$$\alpha^{+}(s)=s^2,\quad \alpha^{-}(s)=(s+2),$$

$$\ker\alpha^{+}(A)=\ker A^2=\operatorname{span}\left\{\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\}.$$

另一方面，

$$\mathscr{B}=\operatorname{Im}B=\operatorname{span}\left\{\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\},$$

$$A\mathscr{B}=\operatorname{span}\left\{\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}\right\},$$

$$A^2\mathscr{B}=A^3\mathscr{B}=0,$$

$$\langle A\mid\mathscr{B}\rangle=\mathscr{B}+A\mathscr{B}+A^2\mathscr{B}+A^3\mathscr{B}$$

$$= \text{span}\left\{\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\}.$$

由此知，

$$\ker \alpha^{+}(\boldsymbol{A}) \subset \langle \boldsymbol{A} \mid \mathscr{B} \rangle.$$

根据定理 14，$(\boldsymbol{A},\boldsymbol{B})$能稳.

2.3.2 能检测性

用有序矩阵对$(\boldsymbol{C},\boldsymbol{A})$来表示系统(2-17)式.

定义 7 称$(\boldsymbol{C},\boldsymbol{A})$能检测(具有**能检测性**)，如果$(\boldsymbol{A}^{\mathrm{T}},\boldsymbol{C}^{\mathrm{T}})$能稳.

定理 15 $(\boldsymbol{C},\boldsymbol{A})$能检测的充分必要条件为

$$\text{rank}\begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} \\ \boldsymbol{C} \end{bmatrix} = n \qquad (s \in \sigma(\boldsymbol{A}) \cap C^{+}), \tag{2-31}$$

其中 $\boldsymbol{I}_n$ 为 n 阶单位矩阵，$\sigma(\boldsymbol{A})$为 $\boldsymbol{A}$ 的谱集，C^{+} 为闭的右半复平面.

定理 16 $(\boldsymbol{C},\boldsymbol{A})$能检测的充分必要条件为

$$\bigcap_{i=0}^{n-1} \ker(\boldsymbol{C}\boldsymbol{A}^{i}) \subset \ker \alpha^{-}(\boldsymbol{A}). \tag{2-32}$$

其中 α^{-}的含义见定理 14 之前的说明.

2.4 定常线性系统的标准结构

设有定常线性系统

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \\ \boldsymbol{y} = \boldsymbol{C}\boldsymbol{x}, \end{cases} \tag{2-33}$$

并用三元组$(\boldsymbol{C},\boldsymbol{A},\boldsymbol{B})$表示(2-33)式.

定义 8 称线性系统$(\boldsymbol{C},\boldsymbol{A},\boldsymbol{B})$是完全的，如果它是能控和能观测的.

定理 17 设系统$(\boldsymbol{C},\boldsymbol{A},\boldsymbol{B})$是完全的，则存在一个坐标变换

$$\boldsymbol{x} = \boldsymbol{T}\bar{\boldsymbol{x}},$$

其中 $\boldsymbol{T}$ 为 $n \times n$ 非奇异矩阵，使得系统$(\boldsymbol{C},\boldsymbol{A},\boldsymbol{B})$在新坐标下具有如下标准结构：

$$\begin{bmatrix} \dot{\bar{\boldsymbol{x}}}_1 \\ \dot{\bar{\boldsymbol{x}}}_2 \\ \dot{\bar{\boldsymbol{x}}}_3 \\ \dot{\bar{\boldsymbol{x}}}_4 \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{A}}_{11} & \boldsymbol{0} & \bar{\boldsymbol{A}}_{13} & \boldsymbol{0} \\ \bar{\boldsymbol{A}}_{21} & \bar{\boldsymbol{A}}_{22} & \bar{\boldsymbol{A}}_{23} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \bar{\boldsymbol{A}}_{33} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \bar{\boldsymbol{A}}_{43} & \bar{\boldsymbol{A}}_{44} \end{bmatrix} \cdot \begin{bmatrix} \bar{\boldsymbol{x}}_1 \\ \bar{\boldsymbol{x}}_2 \\ \bar{\boldsymbol{x}}_3 \\ \bar{\boldsymbol{x}}_4 \end{bmatrix} + \begin{bmatrix} \bar{\boldsymbol{B}}_1 \\ \bar{\boldsymbol{B}}_2 \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{u}, \tag{2-34}$$

$$y=[\bar{C}_1,0,\bar{C}_3,0]\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \bar{x}_3\\ \bar{x}_4\end{bmatrix}.$$

其中 $\bar{x}_1,\bar{x}_2,\bar{x}_3,\bar{x}_4$ 分别是 n_1,n_2,n_3,n_4 维向量，$n_1+n_2+n_3+n_4=n$；$\left(\begin{bmatrix}\bar{A}_{11} & 0\\ \bar{A}_{21} & \bar{A}_{22}\end{bmatrix},\begin{bmatrix}\bar{B}_1\\ \bar{B}_2\end{bmatrix}\right)$是能控对；$\left([\bar{C}_1,\bar{C}_3],\begin{bmatrix}\bar{A}_{11} & \bar{A}_{13}\\ 0 & \bar{A}_{33}\end{bmatrix}\right)$是能观测对；三元组$(\bar{C}_1,\bar{A}_{11},\bar{B}_1)$是完全的.

习惯上，称系统(2-34)式中的 $\bar{x}_1$ 为能控且能观测状态，$\bar{x}_2$ 为能控但不能观测状态，$\bar{x}_3$ 为不能控但能观测状态，$\bar{x}_4$ 为不能控不能观测状态.

标准结构(2-34)式的求法见文献[2].

3　定常线性系统的规范形与实现

3.1　单输入单输出系统的规范形

本节介绍定常系统

$$\begin{cases}\dot{x}=Ax+bu,\\ y=cx\end{cases}\tag{3-1}$$

的规范形. 其中 $x\in\mathbf{R}^n,u\in\mathbf{R},y\in\mathbf{R};A\in\mathbf{R}^{n\times n},b\in\mathbf{R}^{n\times 1},c\in\mathbf{R}^{1\times n}$.

3.1.1　能控规范形

定理 1　设系统(3-1)式能控，A 的特征多项式为

$$\det(sI_n-A)=s^n+a_ns^{n-1}+\cdots+a_2s+a_1,$$

则存在坐标变换

$$x=T\bar{x},$$

其中 T 为 $n\times n$ 非奇异矩阵，使得系统(3-1)式在新坐标系下具有如下规范形(称为能控规范形)：

$$\begin{cases}\dot{\bar{x}}=\bar{A}\,\bar{x}+\bar{b}u;\\ y=\bar{c}\,\bar{x}.\end{cases}\tag{3-2}$$

其中 $\bar{A}=T^{-1}AT,\bar{b}=T^{-1}b,\bar{c}=cT$，并且

$$\bar{A}=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ -a_1 & -a_2 & -a_3 & \cdots & -a_n\end{bmatrix},\quad \bar{b}=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix},$$

$$\bar{c}=[\bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n].$$

能控规范形(3-2)式在系统的极点配置等方面有重要应用.在实际应用中,关键是如何求得能控规范形.下面给出一种求能控规范形的步骤.

(1)计算 A 的特征多项式

$$\det(sI_n-A)=s^n+a_ns^{n-1}+\cdots+a_2s+a_1.$$

(2)构造坐标变换矩阵

$$T=[A^{n-1}b,A^{n-2}b,\cdots,Ab,b]\begin{bmatrix}1&0&\cdots&0&0\\a_n&1&\cdots&0&0\\a_{n-1}&a_n&\cdots&\cdots&\cdots\\\vdots&\vdots&&\vdots&\vdots\\a_2&a_3&\cdots&a_n&1\end{bmatrix}.$$

(3)计算

$$\bar{c}=cT=[\bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n].$$

(4)写出规范形(3-2)式.

注:在上面能控规范形的求法中,$\bar{A}$ 由 A 的特征多项式的系数确定,$\bar{b}$ 是唯一确定的,由步骤(2)和(3)即可求得 $\bar{c}$.

例1　求系统

$$\begin{cases}\dot{x}=\begin{bmatrix}1&0&1\\0&1&0\\1&0&0\end{bmatrix}x+\begin{bmatrix}0\\1\\1\end{bmatrix}u,\\y=[1,1,0]x=cx\end{cases}\tag{3-3}$$

的能控规范形.

解　A 的特征多项式为

$$\det(sI_3-A)=\det\begin{bmatrix}s&-1&-1\\0&s-1&0\\-1&0&s\end{bmatrix}=s^3-2s^2+1.$$

$$Ab=\begin{bmatrix}1&0&1\\0&1&0\\1&0&0\end{bmatrix}\begin{bmatrix}0\\1\\1\end{bmatrix}=\begin{bmatrix}1\\1\\0\end{bmatrix},$$

$$A^2b=\begin{bmatrix}1&0&1\\0&1&0\\1&0&0\end{bmatrix}\begin{bmatrix}1\\1\\0\end{bmatrix}=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

构造变换矩阵

$$T=\begin{bmatrix}1&1&0\\1&1&1\\1&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\-2&1&0\\0&-2&1\end{bmatrix}=\begin{bmatrix}-1&1&0\\-1&-1&1\\1&-2&1\end{bmatrix}.$$

因此,

$$\bar{c}=cT=[1,1,0]\begin{bmatrix}-1&1&0\\-1&-1&1\\1&-2&1\end{bmatrix}=[-2,0,1].$$

系统(3-3)式的能控规范形为

$$\dot{\bar{x}}=\begin{bmatrix}0&1&0\\0&0&1\\-1&0&2\end{bmatrix}\bar{x}+\begin{bmatrix}0\\0\\1\end{bmatrix}u,$$
$$y=[-2,0,1]\bar{x}.$$

3.1.2　能观测规范形

定理 2　设系统(3-1)式能观测.设 A 的特征多项式为

$$\det(sI_n-A)=s^n+a_ns^{n-1}+\cdots+a_2s+a_1,$$

则存在一个坐标

$$x=P\bar{x},$$

其中 P 为 $n\times n$ 非奇异矩阵,使得在新坐标系下(3-1)式具有如下**能观测规范形**:

$$\begin{cases}\dot{\bar{x}}=\bar{A}\bar{x}+\bar{b}u,\\ y=\bar{c}\bar{x}.\end{cases}\tag{3-4}$$

其中 $\bar{A}=P^{-1}AP,\bar{b}=P^{-1}b,\bar{c}=cP$,并且

$$\bar{A}=\begin{bmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&0&0&\cdots&1\\-a_1&-a_2&-a_3&\cdots&-a_n\end{bmatrix},\quad \bar{b}=\begin{bmatrix}b_n\\b_{n-1}\\\vdots\\b_1\end{bmatrix},$$
$$\bar{c}=[1,0,\cdots,0].$$

注:能观测规范形(3-4)式中的 $\bar{A}$ 与能控规范形(3-2)式中的 $\bar{A}$ 是相同的①.

能观测规范形(3-4)式可按下述步骤求得

(1)计算 A 的特征多项式

$$\det(sI_n-A)=s^n+a_ns^{n-1}+\cdots+a_2s+a_1;$$

(2)计算变换矩阵 P 的逆矩阵

$$P^{-1}=\begin{bmatrix}c\\cA\\\vdots\\cA^{n-1}\end{bmatrix};$$

(3)计算

---

①　在计算能观测规范形时,没有用到 P,只用到 P^{-1},而且 P^{-1}可直接计算.

$$\bar{b} = P^{-1}b = \begin{bmatrix} cb \\ cAb \\ \vdots \\ cA^{n-1}b \end{bmatrix};$$

(4)写出能观测规范形(3-4)式.

例 2 求例1中系统(3-3)式的能观测规范形.

解 $\bar{A}$ 与例 1 中的相同,只须计算 $\bar{b}$.注意到:

$$cA = [1,1,0]\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = [1,1,1],$$

$$cA^2 = [1,1,1]\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = [2,1,1],$$

即可得到变换矩阵 P 的逆和 $\bar{b}$:

$$P^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix},$$

$$\bar{b} = P^{-1}b = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

系统(3-3)式的能观测规范形为

$$\dot{\bar{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 2 \end{bmatrix}\bar{x} + \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}u,$$

$$y = [1,0,0]\bar{x}.$$

3.2 多输入多输出系统的规范形

考虑多输入多输出定常系统

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx, \end{cases} \tag{3-5}$$

其中 $x \in \mathbf{R}^n, u \in \mathbf{R}^m, y \in \mathbf{R}^p, A \in \mathbf{R}^{n\times n}, B \in \mathbf{R}^{n\times m}, C \in \mathbf{R}^{p\times n}$.

本节中假设 B 列满秩和 C 行满秩,即 $\mathrm{rank}B = m, \mathrm{rank}C = p$.

3.2.1 能控规范形

定义 1 设(A,B)能控,$\mathscr{B} = \mathrm{Im}B$,记对于 $j = 0,1,\cdots,n-1$,

$$\varphi_j = \mathscr{B} + A\mathscr{B} + A^2\mathscr{B} + \cdots + A^j\mathscr{B} \tag{3-6}$$

令

$$\rho_0 = m,$$
$$\rho_j = d\left(\frac{\varphi_j}{\varphi_{j-1}}\right) \quad (j = 1,2,\cdots,n-1), \tag{3-7}$$

其中 $d\left(\frac{\varphi_j}{\varphi_{j-1}}\right)$ 表示商空间 $\frac{\varphi_j}{\varphi_{j-1}}$ 的维数. 若对于 $i = 1,2,\cdots,m$,

k_i = 集合 $\{\rho_0,\rho_1,\cdots,\rho_{n-1}\}$ 中大于 i 的整数的个数,

整数 $k_1,k_2,\cdots,k_m$ 具有如下性质:

1° $k_1 \geqslant k_2 \geqslant \cdots \geqslant k_m \geqslant 1$.

2° $k_1 + k_2 + \cdots + k_m = n$.

3° 有序数组 $(k_1,k_2,\cdots,k_m)$ 在坐标变换下是不变的, 则 $(k_1,k_2,\cdots,k_m)$ 称为 $(\boldsymbol{A},\boldsymbol{B})$(或系统(3-5)式)的能控性结构指标.

上述能控性结构指标是在数学上定义的, 而对实际应用来说, 只要能实际求出能控性结构指标就可以了. 下面是其一种简单的做法.

设 $\boldsymbol{B} = [\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m]$. 列出向量组

$$\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m;\boldsymbol{A}\boldsymbol{b}_1,\boldsymbol{A}\boldsymbol{b}_2,\cdots,\boldsymbol{A}\boldsymbol{b}_m;\cdots;\boldsymbol{A}^{n-1}\boldsymbol{b}_1,\boldsymbol{A}^{n-1}\boldsymbol{b}_2,\cdots,\boldsymbol{A}^{n-1}\boldsymbol{b}_m. \tag{3-8}$$

然后, 从左到右将每个与其前面的向量线性相关的向量去掉. 经适当调整留下向量的顺序, 可得 n 个线性无关向量的向量组

$$\boldsymbol{b}_1,\boldsymbol{A}\boldsymbol{b}_1,\cdots,\boldsymbol{A}^{r_1-1}\boldsymbol{b}_1;\boldsymbol{b}_2,\boldsymbol{A}\boldsymbol{b}_2,\cdots,\boldsymbol{A}^{r_2-1}\boldsymbol{b}_2;\cdots;\boldsymbol{b}_m,\boldsymbol{A}\boldsymbol{b}_m,\cdots,\boldsymbol{A}^{r_m-1}\boldsymbol{b}_m. \tag{3-9}$$

(3-9)式中的向量组构成 $\mathbf{R}^n$ 中一组基.

(3-9)式中的 $r_1,r_2,\cdots,r_m$ 依大到小排列后, 即为 $(\boldsymbol{A},\boldsymbol{B})$ 的能控结构指标. 因此, 有的文献中, 也将 $\{r_1,r_2,\cdots,r_m\}$ 称为 $(\boldsymbol{A},\boldsymbol{B})$ 的能控性结构指标. 不过应该指出, $\{r_1,r_2,\cdots,r_m\}$ 没有顺序关系, 它不是坐标变换的不变量.

下面, 以 $\{r_1,r_2,\cdots,r_m\}$ 为能控性结构指标(这不改变问题的性质), 给出系统(3-5)式的能控规范形.

定理 3 设系统(3-5)式能控, $(\boldsymbol{A},\boldsymbol{B})$ 的能控性结构指标为 $\{r_1,r_2,\cdots,r_m\}$, 则存在坐标变换

$$\boldsymbol{x} = \boldsymbol{T}\bar{\boldsymbol{x}}, \quad \boldsymbol{u} = \boldsymbol{G}\bar{\boldsymbol{u}},$$

其中 $\boldsymbol{T}$ 为 $n\times n$ 非奇异矩阵, 使得系统(3-5)式在新坐标系下具有如下能控规范形:

$$\begin{cases} \dot{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{u}}, \\ \boldsymbol{y} = \bar{\boldsymbol{C}}\bar{\boldsymbol{x}}. \end{cases} \tag{3-10}$$

其中 $\bar{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}, \bar{\boldsymbol{B}} = \boldsymbol{T}^{-1}\boldsymbol{B}\boldsymbol{G}, \bar{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{T}$, 并且

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \bar{\boldsymbol{A}}_{11} & \bar{\boldsymbol{A}}_{12} & \cdots & \bar{\boldsymbol{A}}_{1m} \\ \bar{\boldsymbol{A}}_{21} & \bar{\boldsymbol{A}}_{22} & \cdots & \bar{\boldsymbol{A}}_{2m} \\ \vdots & \vdots & & \vdots \\ \bar{\boldsymbol{A}}_{m1} & \bar{\boldsymbol{A}}_{m2} & \cdots & \bar{\boldsymbol{A}}_{mm} \end{bmatrix}, \quad \bar{\boldsymbol{B}} = \mathrm{diag}\{\bar{\boldsymbol{b}}_1,\bar{\boldsymbol{b}}_2,\cdots,\bar{\boldsymbol{b}}_m\},$$

$$\bar{A}_{ii}=\begin{bmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&0&0&\cdots&1\\ *&*&*&&*\end{bmatrix}_{r_i\times r_i}\quad (i=1,2,\cdots,m),\tag{3-11}$$

$$\bar{A}_{ij}=\begin{bmatrix}0&0&\cdots&0\\\vdots&\vdots&&\vdots\\0&0&\cdots&0\\ *&*&\cdots&*\end{bmatrix}_{r_i\times r_j}\quad (i\neq j),\tag{3-12}$$

$$\bar{b}_i=\begin{bmatrix}0\\\vdots\\0\\1\end{bmatrix}_{r_i\times 1}\quad (i=1,2,\cdots,m).$$

(3-11)式和(3-12)式中的"$*$"表示该处是某个常数. $\bar{C}$ 没有特殊的形式.

下面给出一种求能控规范形(3-10)式的方法,其步骤如下:

(1) 以(3-9)式中的向量构造矩阵

$Q=[b_1,Ab_1,\cdots,A^{r_1-1}b_1,b_2,Ab_2,\cdots,A^{r_2-1}b_2,\cdots,b_m,Ab_m,\cdots,A^{r_m-1}b_m]$.

(2) 计算 Q^{-1}. 设 q_i^{T} 是 Q^{-1} 中第 $\sum_{j=1}^{i}r_j$ 行的行向量, $i=1,2,\cdots,m$. 构造 T 的逆矩阵 T^{-1}:

$$T^{-1}=\begin{bmatrix}q_1^{\mathrm{T}}\\q_1^{\mathrm{T}}A\\\vdots\\q_1^{\mathrm{T}}A^{r_1-1}\\\vdots\\q_m^{\mathrm{T}}\\q_m^{\mathrm{T}}A\\\vdots\\q_m^{\mathrm{T}}A^{r_m-1}\end{bmatrix}\tag{3-13}$$

(3) 计算 $\bar{A}=T^{-1}AT,\bar{C}=CT$,即可得能控规范形(3-10)式.

下面考虑系统(3-5)式的能观测规范形. 类似于能控性结构指标 $\{r_1,r_2,\cdots,r_m\}$,可以找到一组正整数 $\{s_1,s_2,\cdots,s_p\}$,使得

$$c_1^{\mathrm{T}},(c_1A)^{\mathrm{T}},\cdots,(c_1A^{s_1-1})^{\mathrm{T}},\cdots,c_p^{\mathrm{T}},(c_pA^{s_p-1})^{\mathrm{T}},\cdots,(c_pA^{s_p-1})^{\mathrm{T}}$$

是 $\mathbf{R}^n$ 的一组基,其中 $c_i(i=1,2,\cdots,p)$是 C 的第 i 行.

定义 $\{s_1,s_2,\cdots,s_p\}$ 为系统(3-5)式(C,A)的能观测性结构指标. 自然,也可仿照能控性结构指标 $\{k_1,k_2,\cdots,k_m\}$ 用商空间维数定义能观测性结构指标. 这样定义的指标与坐标选取无关.

定理 4 设系统(3-5)式能观测，能观测结构指标为$\{s_1, s_2, \cdots, s_p\}$，则存在坐标变换

$$x = P\bar{x},$$

其中 P 为 $n \times n$ 非奇异矩阵，使得系统(3-5)式在新坐标系下具有如下能观测规范形：

$$\begin{cases} \dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u, \\ y = \bar{C}\bar{x}, \end{cases} \tag{3-14}$$

其中 $\bar{A} = P^{-1}AP, \bar{B} = P^{-1}B, \bar{C} = CP$，并且

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} & \cdots & \bar{A}_{1p} \\ \bar{A}_{21} & \bar{A}_{22} & \cdots & \bar{A}_{2p} \\ \vdots & \vdots & & \vdots \\ \bar{A}_{p1} & \bar{A}_{p2} & \cdots & \bar{A}_{pp} \end{bmatrix}, \tag{3-15}$$

$$\bar{C} = \mathrm{diag}\{\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_p\},$$

$$\bar{A}_{ii} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ * & * & * & \cdots & * \end{bmatrix}_{s_i \times s_i},$$

$$\bar{A}_{ij} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ * & * & * & \cdots & * \end{bmatrix} \quad (i \neq j), \tag{3-16}$$

$$\bar{c}_i = [0\,0\,0\,\cdots\,0\,1]_{1 \times s_i} \quad (i = 1, 2, \cdots, p),$$

$\bar{B}$ 没有特殊的形式，(3-16)式中的“*”表示该处为某个常数。

能观测规范形的计算步骤如下：

(1) 求出能观测性结构指标$\{s_1, s_2, \cdots, s_p\}$(可用类似于求能控性结构指标$\{r_1, r_2, \cdots, r_m\}$的方法)。

(2) 构造状态变换矩阵 P 的逆矩阵 P^{-1}：

$$P^{-1} = \begin{bmatrix} c_1 \\ c_1 A \\ c_1 A^{s_1 - 1} \\ \vdots \\ c_p \\ c_p A \\ \vdots \\ c_p A^{s_p - 1} \end{bmatrix}.$$

(3) 计算 $\bar{A}=P^{-1}AP$, $\bar{B}=P^{-1}B$, 即得能观测规范形(3-14)式.

3.3 块三角形规范形

考虑定常线性系统(3-5)式. 设(A,B)能控. 记 $B=[b_1,b_2,\cdots,b_m]$, $\mathrm{rank}B=m$, 由能控性理论可知, 存在正整数 q, $1\leqslant q\leqslant m$, 和正整数组$\{\mu_1,\mu_2,\cdots,\mu_q\}$, 并且具有下列性质:

1° $\mu_1+\mu_2+\cdots+\mu_q=n$,

2° $\mu_i\geqslant 1, i=1,2,\cdots,q$,

3° $\mu_i, i=1,2,\cdots,q$, 是使向量组

$$b_1,Ab_1,\cdots,A^{\mu_1-1}b_1,b_2,Ab_2,\cdots,A^{\mu_2-1}b_2,\cdots,b_i,Ab_i,\cdots,A^{\mu_i-1}b_i$$

线性无关的最大正整数.

定理 5 设(A,B)能控, 且正整数组$\{\mu_1,\mu_2,\cdots,\mu_q\}$满足上述性质 1°, 2°, 3°, 则存在坐标变换:

$$x=T\bar{x},$$

其中 T 为 $n\times n$ 非奇异矩阵, 使得系统(3-5)式在新坐标系下具有如下块三角形规范形:

$$\begin{cases}\dot{\bar{x}}=\bar{A}\bar{x}+\bar{B}u,\\ y=\bar{C}\bar{x}.\end{cases}\tag{3-17}$$

其中 $\bar{A}=T^{-1}AT$, $\bar{B}=T^{-1}B$, $\bar{C}=CT$,

$$\bar{A}=\begin{bmatrix}\bar{A}_{11} & 0 & 0 & \cdots & 0\\ \bar{A}_{21} & \bar{A}_{22} & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ \bar{A}_{q1} & \bar{A}_{q2} & \bar{A}_{q3} & \cdots & \bar{A}_{qq}\end{bmatrix},$$

$$\bar{B}=[\bar{b}_1,\ \bar{b}_2,\ \cdots,\ \bar{b}_q,\ \bar{b}_{q+1},\ \cdots,\ \bar{b}_m],$$

$$\bar{A}_{ii}=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ * & * & * & \cdots & *\end{bmatrix}_{\mu_i\times\mu_i}\quad (i=1,2,\cdots,q),\tag{3-18}$$

$$\bar{A}_{ij}=\begin{bmatrix}0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 0\\ * & * & * & \cdots & *\end{bmatrix}_{\mu_i\times\mu_j}\quad (i\neq j),\tag{3-19}$$

$$\bar{b}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{第} \sum_{j=1}^{i} \mu_j \text{行} \quad (i = 1, 2, \cdots, q).$$

(3-18)式和(3-19)式中的“ * ”表示该处是一个常数. $\bar{C} = CT$ 和 $\bar{b}_l$, $l = q+1, \cdots, m$ 没有特殊形式.

类似于能观测范形,亦可得基于能观测假定的块三角形规范形.

3.4　定常线性系统的实现

3.4.1　实现问题的提法

实现问题的提法:已知有理分式矩阵 $G(s)$,求满足

$$G(s) = C(sI_n - A)^{-1}B + D \tag{3-20}$$

的矩阵 A, B, C, D.

定义 2　如果存在(3-20)式的解 A, B, C, D 矩阵,则由矩阵 A, B, C, D 确定的定常线性系统

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx + Du, \end{cases} \tag{3-21}$$

称为 $G(s)$ 的一个状态空间实现,简称实现. 矩阵 A 的阶数称为 $G(s)$ 的实现的阶数.

总用 (C, A, B, D) 或 (C, A, B)(当 $D = 0$ 时)表示 $G(s)$ 的一个实现.

实现问题是系统理论中一个重要问题,它是联系线性系统的频域理论和状态空间理论的一个桥梁.

3.4.2　实现问题的可解性

定义 3　设 $G(s) = [g_{ij}(s)]_{p \times m}$ 为有理分式矩阵. 称 $G(s)$ 为真有理分式矩阵,如果它的每个元 $g_{ij}(s)$, $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, m$, 为真有理分式. 称 $G(s)$ 为严格真有理分式矩阵,如果它的每个元 $g_{ij}(s)$ 为严格真有理分式, $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, m$.

设 $G(s) = [g_{ij}(s)]_{p \times m}$ 为真有理分式矩阵, $d(s)$ 为 $g_{ij}(s)$, $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, m$ 的最小公分母,并具有如下形式:

$$d(s) = s^l + a_l s^{l-1} + \cdots + a_2 s + a_1, \tag{3-22}$$

其中 a_i, $i = 1, 2, \cdots, l$, 为常数. 那么, $G(s)$ 能表示成如下形式:

$$G(s) = \frac{1}{d(s)}(P_l s^{l-1} + P_{l-1} s^{l-2} + \cdots + P_2 + P_1) + Q, \tag{3-23}$$

其中 $P_i, i=1,2,\cdots,l$ 和 Q 为 $p\times m$ 常数矩阵.

由(3-22)式和(3-23)式,可得到 $G(s)$ 的两个实现:能控形实现(C,A,B,D)和能观测形实现$(\bar{C},\bar{A},\bar{B},\bar{D})$,其中

$$A=\begin{bmatrix} 0 & I_m & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & I_m \\ -a_1 I_m & -a_2 I_m & -a_3 I_m & \cdots & -a_l I_m \end{bmatrix}_{lm\times lm}, \tag{3-24}$$

$$B=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ I_m \end{bmatrix}_{lm\times m}, \tag{3-25}$$

$$C=[P_1,\ P_2,\ \cdots,\ P_l]_{p\times lm}, \tag{3-26}$$

$$D=Q, \tag{3-27}$$

其中 I_m 为 m 阶单位矩阵;

$$\bar{A}=\begin{bmatrix} 0 & 0 & \cdots & 0 & -a_1 I_p \\ I_p & 0 & \cdots & 0 & -a_2 I_p \\ 0 & I_p & \cdots & 0 & -a_3 I_p \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & I_p & -a_l I_p \end{bmatrix}_{lp\times lp}, \tag{3-28}$$

$$\bar{B}=\begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_l \end{bmatrix}_{lp\times m}, \tag{3-29}$$

$$\bar{C}=[0,\ 0,\ \cdots,\ 0,\ I_p]_{p\times lp}, \tag{3-30}$$

$$\bar{D}=Q, \tag{3-31}$$

其中 I_p 为 p 阶单位矩阵.

在实现(3-24)式~(3-27)式中,(A,B)是能控对,故称为能控形实现.在(3-28)式~(3-31)式中$(\bar{C},\bar{A})$是能观测对,故称为能观测形实现.

因此真有理分式总存在实现.由此立刻可得如下定理.

定理 6 (实现问题的可解性) 有理分式矩阵 $G(s)$ 存在实现(C,A,B,D)的充分必要条件是其为真有理分式矩阵;存在(C,A,B)实现的充分必要条件是其为严格真有理分式矩阵.

3.4.3 最小实现及其唯一性

在 $G(s)$ 的所有实现中,状态空间维数最小的实现叫做**最小实现**.

定理7　设 $G(s)$ 是严格真有理分式矩阵，则 (C,A,B) 是其最小实现的充分必要条件是 (A,B) 能控和 (C,A) 能观测.

定理8　$G(s)$ 的最小实现在代数等价意义下是唯一的，即如果 (C,A,B) 和 $(\bar{C},\bar{A},\bar{B})$ 都是 $G(s)$ 的最小实现，则 (C,A,B) 和 $(\bar{C},\bar{A},\bar{B})$ 是代数等价的.

3.4.4　最小实现的计算

最小实现有多种计算方法，下面介绍的是一种概念上较简单的方法，其他的计算方法见文献[2].

设 $G(s)$ 是严格真有理分式矩阵，其最小实现的具体计算步骤如下：

(1) 求 $G(s)$ 各元的最小公分母 $d(s)$，并将其表示成(3-22)式的形式.

(2) 将 $G(s)$ 展开成(3-23)式的形式.

(3) 构造 $G(s)$ 的能控形实现 (C,A,B)（或能观测形实现 $(\bar{C},\bar{A},\bar{B})$）.

(4) 根据第2章的结构分解方法，从 (C,A,B)（或 $(\bar{C},\bar{A},\bar{B})$）中分出能观测部分（或能控部分），即为 $G(s)$ 的最小实现.

例3　求

$$G(s)=\begin{bmatrix}\dfrac{1}{s(s+1)} & \dfrac{1}{s+2}\\ \dfrac{1}{s+1} & \dfrac{1}{s+1}\end{bmatrix}$$

的最小实现.

解　$G(s)$ 的最小公分母为

$$d(s)=s^3+3s^2+2s.$$

$$G(s)=\frac{1}{d(s)}\left(s^2\begin{bmatrix}0 & 1\\ 1 & 1\end{bmatrix}+s\begin{bmatrix}1 & 1\\ 2 & 2\end{bmatrix}+\begin{bmatrix}2 & 0\\ 0 & 0\end{bmatrix}\right).$$

作能控形实现 (C,A,B)，其中

$$A=\begin{bmatrix}0 & I_2 & 0\\ 0 & 0 & I_2\\ 0 & -2I_2 & -3I_2\end{bmatrix}_{6\times 6},\quad B=\begin{bmatrix}0\\ 0\\ I_2\end{bmatrix}_{6\times 2},$$

$$C=\begin{bmatrix}2 & 0 & 1 & 1 & 0 & 1\\ 0 & 0 & 2 & 2 & 1 & 1\end{bmatrix}.$$

再求 (C,A,B) 的能观测部分. 为计算状态变换矩阵 T（T 的详细计算方法见第2章），经计算得

$$\bigcap_{i=0}^{5}\ker(CA^i)=\mathrm{span}\left\{\begin{bmatrix}0\\ 1\\ 0\\ 0\\ 0\\ 0\end{bmatrix}\right\},$$

于是，取变换矩阵

$$T=\begin{bmatrix}1&0&0&0&0&0\\0&0&0&0&0&1\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\end{bmatrix}.$$

经计算有

$$\bar{A}=T^{-1}AT=\begin{bmatrix}0&1&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&-2&0&-3&0&0\\0&0&-2&0&-3&0\\0&0&1&0&0&0\end{bmatrix},$$

$$\bar{B}=T^{-1}B=\begin{bmatrix}0&0\\0&0\\0&0\\1&0\\0&1\\0&0\end{bmatrix},$$

$$\bar{C}=CT=\begin{bmatrix}2&2&1&0&1&0\\0&2&2&1&1&0\end{bmatrix}.$$

由此易得 $G(s)$的最小实现(C_1,A_1,B_1)为

$$A_1=\begin{bmatrix}0&1&0&0&0\\0&0&0&1&0\\0&0&0&0&1\\0&-2&0&-3&0\\0&0&-2&0&-3\end{bmatrix},\quad B_1=\begin{bmatrix}0&0\\0&0\\0&0\\1&0\\0&1\end{bmatrix},$$

$$C_1=\begin{bmatrix}2&2&1&0&1\\0&2&2&1&1\end{bmatrix}.$$

4 极点配置和观测器设计

本章介绍定常线性系统

$$\begin{cases}\dot{x}=Ax+Bu,\\ y=Cx\end{cases}\tag{4-1}$$

的极点配置和观测器设计问题,其中 $x\in\mathbf{R}^n,u\in\mathbf{R}^m,y\in\mathbf{R}^p,A\in\mathbf{R}^{n\times n},B\in\mathbf{R}^{n\times m}$, $C\in\mathbf{R}^{p\times n}$.

4.1　状态反馈极点配置

4.1.1　问题提法

设 Λ 是 l 个复数的集合,如果 Λ 具有形式

$$\Lambda = \{\alpha_1, \alpha_2, \cdots, \alpha_{l_1}, \beta_1, \beta_2, \cdots, \beta_{l_2}, \bar{\beta}_1, \bar{\beta}_2, \cdots, \bar{\beta}_{l_2}\}, \tag{4-2}$$

其中 α_i 是实数,$i = 1, 2, \cdots, l_1$,β_j 是非实复数,$j = 1, 2, \cdots, l_2$,$\bar{\beta}_j$ 是β_j 的共轭复数,则称 Λ 是关于实轴的对称集,简称对称集.

定义 1　考虑系统(4-1)式,如果对于任意给定的 n 个复数的对称集Λ,存在 $\boldsymbol{K} \in \mathbf{R}^{m \times n}$,使得矩阵 $\boldsymbol{A} + \boldsymbol{BK}$ 的特征值集合为Λ,则称系统(4-1)式能任意极点配置,或称$(\boldsymbol{A}, \boldsymbol{B})$能任意极点配置.

4.1.2　能任意极点配置的条件

定理 1　$(\boldsymbol{A}, \boldsymbol{B})$能任意极点配置的充分必要条件是$(\boldsymbol{A}, \boldsymbol{B})$能控.

4.1.3　极点配置的算法

设$(\boldsymbol{A}, \boldsymbol{B})$能控,$\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$为给定的对称集.极点配置问题指求 $\boldsymbol{K} \in \mathbf{R}^{m \times n}$,使得 $\boldsymbol{A} + \boldsymbol{BK}$ 的特征值集合为Λ.下面给出 $\boldsymbol{K}$ 的求法.

(1) **单输入系统 $\boldsymbol{K}$ 的求法**　设 $m = 1$,$\boldsymbol{B} = \boldsymbol{b}$.用坐标变换 $\boldsymbol{x} = \boldsymbol{T}\bar{\boldsymbol{x}}$,将$(\boldsymbol{A}, \boldsymbol{b})$化为能控规范形$(\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}})$.

$$\bar{\boldsymbol{A}} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_n \end{bmatrix}, \quad \bar{\boldsymbol{b}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{4-3}$$

闭环系统的特征多项式为

$$\begin{aligned} f(\lambda) &= (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) \\ &\overset{\text{def}}{=\!=} \lambda^n + \beta_n\lambda^{n-1} + \cdots + \beta_2\lambda + \beta_1, \end{aligned} \tag{4-4}$$

其中 β_i 为实数,$i = 1, 2, \cdots, n$.

记

$$\bar{\boldsymbol{k}}^{\mathrm{T}} = [\alpha_1 - \beta_1, \alpha_2 - \beta_2, \cdots, \alpha_n - \beta_n]. \tag{4-5}$$

取

$$\boldsymbol{k}^{\mathrm{T}} = \bar{\boldsymbol{k}}^{\mathrm{T}}\boldsymbol{T}^{-1}, \tag{4-6}$$

则 $\boldsymbol{K} = \boldsymbol{k}^{\mathrm{T}}$ 即为所求.

(2) **多输入系统 $\boldsymbol{K}$ 的求法**　主要思想是,将问题化为单输入问题,然后用(1)的方法求解.

不失一般性,设 $\operatorname{rank}\boldsymbol{B} = m$,$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m]$.

求出 $\boldsymbol{F} \in \mathbf{R}^{m \times n}$,使得$(\boldsymbol{A} + \boldsymbol{BF}, \boldsymbol{b}_1)$为能控对.$\boldsymbol{F}$ 的求法见下面(3).

利用第(1)点,求出 $\boldsymbol{k}_1^{\mathrm{T}} \in \mathbf{R}^{1 \times n}$,使得 $\boldsymbol{A} + \boldsymbol{BF} + \boldsymbol{b}_1\boldsymbol{k}_1^{\mathrm{T}}$ 的特征值集合恰好为 Λ.

取

$$K = F + \bar{b}_1 k_1^T, \tag{4-7}$$

其中 $\bar{b}_1 = [1,0,\cdots,0]^T \in \mathbf{R}^{m\times 1}$. K 即为所求.

(3) F 的求法　设 k_1 是使向量组

$$b_1, Ab_1, \cdots, A^{k_1-1}b_1 \tag{4-8}$$

独立的最大整数. 如果 $k_1 = n$, 则取 $F = 0$. 设 $k_1 < n$, 则

$$x_1 = b_1, \tag{4-9}$$

$$x_i = Ax_{i-1} + b_1 \qquad (i = 2,3,\cdots,k_1),$$

显然, $x_1, x_2, \cdots, x_{k_1}$ 是独立的.

设 k_2 是使向量组

$$x_1, x_2, \cdots, x_{k_1}, b_2, Ab_2, \cdots, A^{k_2-1}b_2$$

为独立的最大整数, 定义

$$x_{k_1+i} = Ax_{n_1+i-1} + b_2 \quad (i = 1,2,\cdots,k_2). \tag{4-10}$$

同样可以证明 $x_1, x_2, \cdots, x_{k_1+k_2}$ 是独立的. 按此进行, 可以找到独立的 n 个独立向量 $x_1, x_2, \cdots, x_n, x_{i+1}$ 具有如下形式:

$$x_{i+1} = Ax_i + b_i \quad (i = 1,2,\cdots,n-1), \tag{4-11}$$

其中 $\tilde{b}_i$ 是 B 的列.

设

$$\tilde{b}_i = B\,\tilde{u}_i \quad (i = 1,2,\cdots,n). \tag{4-12}$$

选择　$F \in \mathbf{R}^{m\times n}$, 使得

$$Fx_i = \tilde{u}_i \quad (i = 1,2,\cdots,n). \tag{4-13}$$

F 即为所求.

例 1　设

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

求 K, 使得

$$\sigma(A + BK) = \{-1, -1, -1+i, -1-i\}.$$

解　令 $b_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$. 先选取 F, 使得 $(A + BF, b_1)$ 能控. F 的选取可按上述第(3)点进行. 对于具体问题, 亦可随机地选取, 这是因为能控性具有通有性. 这里取

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

并记

$$A_F = A + BF = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

不难验证：(A_F, b_1)能控，且 A_F 的特征多项式为

$$\det(\lambda I_4 - A_F) = \lambda^4 - \lambda^3,$$

注意到本题中有 $\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = -1$.

计算状态变换矩阵 T：

$$T = [A_F^3 b_1,\ A_F^2 b_1,\ A_F b_1,\ b_1] \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$T^{-1} = \begin{bmatrix} -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

闭环系统所要求的特征多项式为

$$f(\lambda) = (\lambda+1)(\lambda+1)(\lambda+1-i)(\lambda+1+i) = \lambda^4 + 4\lambda^3 + 7\lambda^2 + 6\lambda + 2.$$

因此，

$$\beta_1 = 2, \beta_2 = 6, \beta_3 = 7, \beta_4 = 4.$$

取

$$\bar{k}^{\mathrm{T}} = [\beta_1 - \alpha_1, \beta_2 - \alpha_2, \beta_3 - \alpha_3, \beta_4 - \alpha_4] = [2,6,7,5].$$

由此，可得

$$K = \bar{b}_1 \bar{k}^{\mathrm{T}} T^{-1} + F$$

$$= \begin{bmatrix} 1 \\ 0 \end{bmatrix} [2,6,7,5] \begin{bmatrix} -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 2 & 6 & 9 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

4.2 动态反馈极点配置

4.2.1 动态输出反馈概念

输出反馈分为静态输出反馈和动态输出反馈.静态输出反馈一般不能任意极点配置,因此,在任意极点配置问题上,常用动态输出反馈.

动态输出反馈的一般形式为

$$\begin{cases} \dot{\xi} = A_c\xi + Ry, \\ u = Q\xi + Hy, \end{cases} \tag{4-14}$$

其中 $\xi \in \mathbf{R}^l$, A_c, R, Q 和 H 分别是 $l \times l, l \times p, m \times l$ 和 $m \times p$ 常值矩阵.

动态输出反馈控制律(4-14)式也叫做系统(4-1)式的一个动态补偿器,ξ 也叫做动态补偿的状态向量.

由系统(4-1)式和(4-14)式组成的闭环系统为

$$\begin{bmatrix} \dot{x} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} A + BHC & BQ \\ RC & A_c \end{bmatrix} \begin{bmatrix} x \\ \xi \end{bmatrix}. \tag{4-15}$$

4.2.2 极点配置

用动态输出反馈配置极点,与动态补偿器(4-14)式的状态向量的维数 l 有关. l 多大才能使闭环系统(4-15)式的 $n+l$ 个极点能任意配置呢? 这里需先引入能控性指标和能观测性指标的概念.

定义 2 设(A, B)能控,$\mathscr{B} = \mathrm{Im}B$.定义$(A, B)$的能控性指标为

$$k_c = \min\{j \mid 1 \leqslant j \leqslant n, \mathscr{B} + A\mathscr{B} + \cdots + A^{j-1}\mathscr{B} = \mathbf{R}^n\}. \tag{4-16}$$

定义 3 定义(C, A)(或系统(4-1)式)的能观测性指标为

$$k_0 = \min\{j \mid 1 \leqslant j \leqslant n, \bigcap_{i=1}^{j} \ker(CA^{i-1}) = 0\}, \tag{4-17}$$

显然 $1 \leqslant k_c \leqslant n, 1 \leqslant k_0 \leqslant n$.

定理 2 设系统(4-1)式能控,且能观测,(C, A)的能观测性指标为 k_0,则对于任意给定的 $n + k_0 - 1$ 个复数的对称集 Λ,都存在一个动态输出反馈控制律(4-14)式,使得闭环系统(4-15)式以 Λ 为极点集,其中 $l = k_0 - 1$.

定理 3 设系统(4-1)式能控,且能观测,(A, B)的能控性指标为 k_c,则对于任意给定的 $n + k_c - 1$ 个复数的对称集 Λ,存在一个动态输出反馈控制律(4-14)式,使得闭环系统(4-15)式,以 Λ 为其极点集,其中 $l = k_c - 1$.

注:当定理 2 和定理 3 中的 $l = 0$ 时,反馈控制律是静输出反馈控制律.

动态补偿器的计算较复杂,限于篇幅这里省去.读者可见参考文献[1]和[2].

4.3 状态观测器设计

4.3.1 观测器概念

在线性控制系统理论中，系统的极点配置、解耦控制及线性二次最优控制等都涉及到状态反馈，但由于状态不易直接量测，或者由于量测在经济上技术上受到的限制，系统状态不总是都能获取的，于是提出了状态重构或状态估计问题. 状态重构问题的实质是重新构造一系统，以原系统的测量和输入作为输入. 如果新系统的输出 z 在某种意义上等价于原系统的状态 $\boldsymbol{x}$，则称 z 是 $\boldsymbol{x}$ 的重构或估计. 用以实现状态重构的新系统称为原系统的观测器.

如果重构的是系统的全部状态，则相应的观测器称为状态观测器；如果重构的是系统状态的一个函数量，则相应的观测器称为函数观测器.

4.3.2 全阶状态观测器

1. 全阶状态观测器概念

给定系统(4-1)式，构造如下形式的动态系统：

$$\dot{z} = \boldsymbol{F}z + \boldsymbol{N}\boldsymbol{u} + \boldsymbol{G}\boldsymbol{y}, \tag{4-18}$$

其中 $z \in \mathbf{R}^n$ 为(4-18)式的状态向量，$\boldsymbol{u}$ 和 $\boldsymbol{y}$ 分别是系统(4-1)式的输入和输出，$\boldsymbol{F} \in \mathbf{R}^{n \times n}$，$\boldsymbol{N} \in \mathbf{R}^{n \times m}$，$\boldsymbol{G} \in \mathbf{R}^{n \times p}$.

令

$$\boldsymbol{e} = z - \boldsymbol{x} \tag{4-19}$$

为系统(4-1)式和(4-18)式的状态误差. 如果对于任意的初始条件 $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ 和 $z(t_0) = z_0$，以及任意的 $\boldsymbol{u}(t)$，皆有

$$\lim_{t \to +\infty} \boldsymbol{e}(t) = 0, \tag{4-20}$$

则称系统(4-18)式为系统(4-1)式的全阶状态观测器，简称为状态观测器.

2. 状态观测器的结构定理和存在性

定理 4 (结构定理)　设系统(4-1)式能控并能观测，则系统(4-18)式是系统(4-1)式的状态观测器的充分必要条件为

1° $\boldsymbol{F}$ 的特征值皆具有负实部；

2° $\boldsymbol{F} = \boldsymbol{A} - \boldsymbol{G}\boldsymbol{C}$；

3° $\boldsymbol{N} = \boldsymbol{B}$.

下面的定理给出了状态观测器存在的条件.

定理 5　系统(4-1)式存在形如系统(4-18)式的状态观测器的充分必要条件为 $(\boldsymbol{C}, \boldsymbol{A})$ 能检测.

3. 构造状态观测器的算法

设 $(\boldsymbol{C}, \boldsymbol{A})$ 能观测. 根据状态观测器的结构定理 4，只须选择 $\boldsymbol{G}$，使 $\boldsymbol{A} - \boldsymbol{G}\boldsymbol{C}$ 的特征值具有负实部. 但是在实际应用中，一般要求误差的收敛比被观测系统(4-1)式

的响应要快，因此，常常要求观测器具有指定的极点．

设 Λ 是观测器被指定的极点对称集，则全阶状态观测器的设计步骤如下：

(1) 导出对偶系统 $(\boldsymbol{B}^{\mathrm{T}},\boldsymbol{A}^{\mathrm{T}},\boldsymbol{C}^{\mathrm{T}})$．

(2) 利用极点配置问题的算法，对矩阵对 $(\boldsymbol{A}^{\mathrm{T}},\boldsymbol{C}^{\mathrm{T}})$ 选取矩阵 $\boldsymbol{K}\in\mathbf{R}^{p\times n}$，使得

$$\sigma(\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{C}^{\mathrm{T}}\boldsymbol{K})=\Lambda.$$

(3) 取 $\boldsymbol{G}=\boldsymbol{K}^{\mathrm{T}}$．

(4) 计算 $\boldsymbol{A}-\boldsymbol{GC}$，则所要设计的状态观测器为

$$\dot{z}=(\boldsymbol{A}-\boldsymbol{GC})z+\boldsymbol{Bu}+\boldsymbol{Gy}. \tag{4-21}$$

4.3.3 最小阶状态观测器

1．最小阶状态观测器概念

状态观测器(4-18)式的动态阶数是 n，与被观测系统(4-1)式的动态阶数相同．但观测器的阶 n 有可能降低，因为若输出映像具有秩 p，则从 $\boldsymbol{y}(t)$ 就能直接算出 $\boldsymbol{x}(t)$ 在 p 维商空间 $\dfrac{\mathbf{R}^n}{\ker\boldsymbol{C}}$ 中的陪集．下面将提出一个动态观测器的构造方法．这个观测器和(4-18)式有同样的形式，它给出了 $\boldsymbol{x}(t)$ 在 $n-p$ 维空间 $\ker\boldsymbol{C}$ 中的“分量”．如果每个状态 $\boldsymbol{x}\in\mathbf{R}^n$ 都从初始状态 $\boldsymbol{x}_0$ 可达，则要求 $z(t)$ 借助于 $\boldsymbol{y}(t)$ 在 $t\to+\infty$ 时的极限情形给出 $\boldsymbol{x}(t)$ 的一种渐近识别，则具有一般形式的(4-18)式的观测器的阶数不可能小于 $n-p$，在这个意义上讲，$n-p$ 阶观测器是“最小阶观测器”．

2．最小阶观测器的存在性

最小阶观测器有如下存在性定理．

定理 6 设 $(\boldsymbol{C},\boldsymbol{A})$ 能观测，$\mathrm{rank}\,\boldsymbol{C}=p$．设 $\Lambda\subset\boldsymbol{C}$ 是对称的，$|\Lambda|=n-p$，则存在一个子空间 $\mathscr{Y}\subset\mathbf{R}^n$，$\dim\mathscr{Y}=n-p$，以及映射 $\boldsymbol{G}:\mathrm{Im}\boldsymbol{C}\to\mathbf{R}^n$，$\boldsymbol{F}:\mathscr{Y}\to\mathscr{Y}$ 和 $\boldsymbol{V}:\mathbf{R}^n\to\mathscr{Y}$，使得

$$\boldsymbol{V}(\boldsymbol{A}-\boldsymbol{GC})=\boldsymbol{FV},\quad \sigma(\boldsymbol{F})=\Lambda,$$

并且

$$\boldsymbol{Q}:\mathbf{R}^n\to\mathrm{Im}\boldsymbol{C}\oplus\mathscr{Y},\quad \boldsymbol{x}\to\boldsymbol{Cx}\oplus V\boldsymbol{x}$$

是同构的．这里 $|\Lambda|$ 表示集合 Λ 元的个数．

假设 $(\boldsymbol{C},\boldsymbol{A})$ 能观测，则根据定理 6，可以构造出如下的最小阶观测器：

$$\dot{z}=\boldsymbol{F}z+\boldsymbol{VGy}+\boldsymbol{VBu}, \tag{4-22}$$

其中 $\boldsymbol{F},\boldsymbol{V},\boldsymbol{G}$ 由定理 6 给出．

系统(4-22)式是系统(4-1)式的观测器的意思是指，将 $\boldsymbol{Q}^{-1}(\boldsymbol{y}(t)\oplus z(t))$ 看做该观测器的输出，则当 $\Lambda\subset C^-$ 时，误差

$$\boldsymbol{e}(t)=\boldsymbol{Q}^{-1}(\boldsymbol{y}(t)\oplus z(t))-\boldsymbol{x}(t), \tag{4-23}$$

当 $t\to+\infty$ 时，以负指数规律趋于零．

实际中，应选择 Λ 使(4-23)式的误差比被观测系统(4-1)式的响应更快地趋于零．

3．最小阶观测器的设计

设 $(\boldsymbol{C},\boldsymbol{A})$ 能观测，Λ 为观测器所期望的极点集，则最小阶观测器的设计步骤

如下.

(1) 将系统(4-1)式化为如下形式:

$$\begin{cases}\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1u,\\ \dot{x}_2 = A_{21}x_1 + A_{22}x_2 + B_2u,\\ y = x_1.\end{cases} \tag{4-24}$$

容易证明:(4-24)式中的(A_{11},A_{12})是能观测对.

(2) 根据极点配置算法,对能控对$(A_{22}^{\mathrm{T}},A_{12}^{\mathrm{T}})$,选取$G^{\mathrm{T}}$,使得

$$\sigma(A_{22}^{\mathrm{T}} - A_{12}^{\mathrm{T}}G^{\mathrm{T}}) = \Lambda.$$

(3) 根据上述数据,构造观测器如下:

$$\begin{gathered}\dot{z} = (A_{22} - GA_{12})z + (B_2 - GB_1)u + \\ [(A_{21} - GA_{11}) + (A_{22} - GA_{12})G]y;\\ W = \begin{bmatrix} y \\ z + Gy\end{bmatrix}.\end{gathered} \tag{4-25}$$

$W(t)$即为$\begin{bmatrix} x_1(t) \\ x_2(t)\end{bmatrix}$的重构.如果要重构$x(t)$,则应用$TW(t)$,其中$T$是将系统(4-1)式变为系统(4-24)式所用的变换$x = T\cdot\begin{bmatrix} x_1 \\ x_2\end{bmatrix}$的变换矩阵.

4.3.4　函数观测器

以重构$Kx(t)$,K为$q\times n$矩阵,为目标的观测器称为系统(4-1)式的函数观测器.

函数观测器的一般形式为

$$\begin{cases}\dot{z} = Fz + Gy + Nu,\\ W = Mz + Hy,\end{cases} \tag{4-26}$$

其中$z\in\mathbf{R}^l, W\in\mathbf{R}^l, F\in\mathbf{R}^{l\times l}, G\in\mathbf{R}^{l\times p}, N\in\mathbf{R}^{l\times m}, M\in\mathbf{R}^{q\times l}, H\in\mathbf{R}^{q\times p}$.

设计函数观测器的目的是,对于任意的初值$z(t_0) = z_0, x(t_0) = x_0$和任意的$u(t)$,都有

$$e(t) = W(t) - KX(t) \to 0 \qquad (t\to+\infty). \tag{4-27}$$

系统(4-26)式成为函数观测器的充分必要条件为:存在$l\times m$矩阵T,使得

1° $TA - FT = GC$;

2° $N = TB$;

3° F的特征值都有负实部;

4° $MT + HC = K$.

设计函数观测器的一个重要问题是如何确定观测器(4-26)式的动态阶数l.这是一个复杂的问题.如果K为$1\times n$矩阵,则(4-26)式称为泛函观测器,其动态阶数l可取为$k_0 - 1$,k_0是(C,A)的能观测指标.有关泛函观测器的设计方法见文献[1].

5 一般线性调节理论

5.1 调节问题的描述

一般线性调节系统的数学模型为

$$\begin{cases} \dot{x}_1 = A_1x_1 + A_3x_2 + B_1u, \\ \dot{x}_2 = A_2x_2, \\ y = C_1x_1 + C_2x_2, \\ z = D_1x_1 + D_2x_2. \end{cases} \tag{5-1}$$

其中 $A_1\in \mathbf{R}^{n_1\times n_1}, A_2\in \mathbf{R}^{n_2\times n_2}, A_3\in \mathbf{R}^{n_1\times n_2}; B_1\in \mathbf{R}^{n_1\times m}; C_1\in \mathbf{R}^{p\times n_1}, C_2\in \mathbf{R}^{p\times n_2}; D_1\in \mathbf{R}^{q\times n_1}, D_2\in \mathbf{R}^{q\times n_2}$；$x_1$ 表示系统的状态变量，$x_1\in \mathbf{R}^{n_1}$；x_2 表示外部输入，可以是干扰输入，也可以是参考信号输入，$x_2\in \mathbf{R}^{n_2}$；u 是控制输入，$u\in \mathbf{R}^m$；y 是量测输出，$y\in \mathbf{R}^p$；z 是被调节输出，$z\in \mathbf{R}^q$.

将量测输出与被调节输出分离的意思是，量测量 y 是设计控制器唯一可以利用的信息，在一般情况下，可以不同于被调节输出 z. 当然，在特殊情况下，它们也可以相同.

如果在(5-1)式中，$C_2=0, D_2=0$，则(5-1)式表示一个纯调节系统；如果 $A_3=0$，则(5-1)式表示一个纯跟踪系统，$z(t)$表示跟踪误差. 一般地说(5-1)式表示带有干扰输入和参考输入的跟踪系统. 因为跟踪问题和调节问题没有本质的区别，因此，系统(5-1)式可看做一个调节系统.

为便于研究，对于系统(5-1)式常作下面的基本假设：

1° $\sigma(A_2)\subset C^+$； (5-2)

2° $\operatorname{rank}B_1=m, \operatorname{rank}[C_1, C_2]=p, \operatorname{rank}D_1=q$. (5-3)

假设 1°的实际意义是：外部输入 $x_2(t)$的稳定部分，会很快趋于零，在实际设计时可以忽略不计，而只须考虑 $x_2(t)$的不趋于零的部分. 假设 2°是一点技术性的限制，不影响问题的性质.

研究调节系统的主要目的是：设计一个动态补偿器

$$\begin{cases} \dot{x}_c = A_cx_c + B_cy, \\ u = F_cx_c + Fy. \end{cases} \tag{5-4}$$

其中 $x_c\in \mathbf{R}^l$，是动态补偿器的状态变量；A_c, B_c, F_c 和 F 分别是 $l\times l, l\times p, m\times l$ 和 $m\times p$ 矩阵，使得闭环系统

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_c \end{bmatrix} = \begin{bmatrix} A_1+B_1FC_1 & B_1F_c \\ B_cC_1 & A_c \end{bmatrix} \begin{bmatrix} x_1 \\ x_c \end{bmatrix} + \begin{bmatrix} A_3+B_1FC_2 \\ B_cC_2 \end{bmatrix} x_2, \tag{5-5}$$

$$\dot{x}_2 = A_2 x_2, \quad z = D_1 x_1 + D_2 x_2,$$

具有如下性质：

1° 它是内部稳定的，即

$$\sigma\left(\begin{bmatrix} A_1 + B_1 F C_1 & B_1 F_c \\ B_c & A_c \end{bmatrix}\right) \subset C^-; \tag{5-6}$$

2° 它是输出调节的，即对于任意 $x_1(t_0) = x_{10}, x_2(t_0) = x_{20}, x_c(t_0) = x_{c0}$，都有

$$\lim_{t \to +\infty} z(t) = \mathbf{0}. \tag{5-7}$$

定义 1 设给定系统(5-1)式，如果存在形如(5-4)式的动态补偿器，使得闭环系统(5-5)式是内部稳定，并且输出调节的，则称调节问题(5-1)式可解，并称动态补偿器(5-4)式是系统(5-1)式的一个综合或无静差补偿器. 带有无静差补偿器的闭环系统叫做无静差系统或叫做稳定的输出调节系统.

5.2 输出调节系统的结构引理

考虑闭环系统(5-5)式，记

$$A_L = \begin{bmatrix} A_1 + B_1 F C_1 & B_1 F_c \\ B_c C_1 & A_c \end{bmatrix}, \quad B_L = \begin{bmatrix} A_3 + B_1 F C_2 \\ B_c C_2 \end{bmatrix},$$

$$D_L = [D_1, \mathbf{0}], \quad x_L = \begin{bmatrix} x_1 \\ x_c \end{bmatrix},$$

则闭环系统(5-5)式，可以写成如下形式：

$$\begin{cases} \dot{x}_L = A_L x_L + B_L x_2, \\ \dot{x}_2 = A_2 x_2, \\ z = D_L x_L + D_2 x_2. \end{cases} \tag{5-8}$$

引理 1(输出调节系统的结构引理) 考虑系统(5-8)式. 设 $\sigma(A_L) \subset C^-$，则系统(5-8)式是输出调节的充分必要条件为，存在 $(n_1 + l) \times n_2$ 矩阵 V，使得下面等式成立：

$$\begin{aligned} A_L V - V A_2 &= B_L, \\ D_L V &= D_2. \end{aligned} \tag{5-9}$$

推论 1 设 $\sigma(A_L) \subset C^-$，则系统(5-8)式是输出调节的充分必要条件为，存在 $n_1 \times n_2$ 矩阵 V_1 和 $l \times n_2$ 矩阵 V_2，使得下面等式成立：

$$\begin{aligned} (A_1 + B_1 F C_1) V_1 - V_1 A_2 + B_1 F_c V_2 &= A_3 + B_1 F C_2, \\ B_c C_1 V_1 + A_c V_2 - V_2 A_2 &= B_c C_2, \\ D V_1 &= D_2. \end{aligned} \tag{5-10}$$

上述引理 1 和推论 1 对研究调节系统的结构很有用.

5.3 带有干扰补偿的动态补偿器

在反馈控制系统理论中，两个最重要的问题是系统的稳定性和抗干扰能力，而

稳定性设计也是为了提高系统的抗干扰能力所采取的技术措施.因此可以说,一个反馈控制系统的好坏主要取决于它的抗干扰能力.本节介绍一种用干扰补偿的方法来消除外部干扰的动态补偿器的设计方法.

5.3.1 干扰补偿的基本思想

研究系统

$$\begin{cases} \dot{x}_1 = A_1x_1 + A_3x_2 + B_1u, \\ \dot{x}_2 = A_2x_2, \\ y = C_1x_1 \end{cases} \tag{5-11}$$

的调节问题,其中以上各式符号的含义与5.1.1小节中相同.

系统(5-11)式是由系统(5-1)式令 $y(t)=z(t), C_2=D_2=0$ 得到的.因此,在系统(5-11)式中, y 既是量测输出,也是被调节的输出.

干扰补偿方法的基本思想是:将控制 $u(t)$ 分成两部分,一部分用于控制系统,使其内部稳定,另一部分用于补偿外部干扰.设

$$u = u_c + u_e, \tag{5-12}$$

其中 u_c 用于控制系统, u_e 用于补偿干扰.假设系统的外部干扰 x_2 可以被量测得到,并且存在一个 $m \times n_2$ 矩阵 E,使得

$$A_3 = B_1E, \tag{5-13}$$

那么,只要取

$$u_e(t) = -Ex_2(t), \tag{5-14}$$

就能消除外部干扰对系统的影响,即外部干扰得到了补偿.而 u_c 可以按前几章所述的方法进行处理.

条件(5-13)式等价于

$$\mathrm{rank}[B_1, A_3] = \mathrm{rank}B_1,$$

即 A_3 的每一列都是 B_1 的列的线性组合.这是干扰能补偿的条件.

除了要有干扰能补偿的条件外,还要求干扰能量测.直接量测干扰 x_2 一般是不可能的,但如果将 x_2 看做状态变量,系统(5-11)式就成了一个复合系统.若将 y 看做整个复合系统的输出,那么在复合系统能观测的条件下,就能设计出一种状态观测器,由它得到干扰的估计.然后用 x_2 的估计 $\hat{x}_2$ 去代替反馈(5-14)式中的 x_2.这样就能做到用动态补偿器实现输出调节变量的目的.

5.3.2 调节问题可解的条件

定理1 考虑系统(5-11)式.设 (A_1, B_1) 能控, (C_1, A_1) 能观测,如果

1° $\mathrm{rank}B_1 = \mathrm{rank}[B_1, A_3]$; (5-15)

2° 对于每个 $\lambda \in \sigma(A_2)$,有

$$\mathrm{rank}\begin{bmatrix} A_1 - \lambda I_{n_1} & A_3 \\ 0 & A_2 - \lambda I_{n_2} \\ C_1 & 0 \end{bmatrix} = n_1 + n_2, \tag{5-16}$$

其中 I_{n_1} 为 $n_1 \times n_1$ 阶单位矩阵，那么，存在一个带干扰补偿的动态补偿器，使得闭环系统内部稳定和输出调节.

进一步研究表明，从定理 1 的条件可知，实际上可以设计一个全状态输出调节器.所谓全状态调节器，就是以 x_1 为被调节输出变量的调节器，即 $y = x_1$.关于全状态调节器有下面的充分必要条件.

定理 2 考虑定常线性系统(5-11)式.它存在全状态输出调节器的充分必要条件为

1° (A_1, B_1)能稳，(C_1, A_1)能检测；

2° $\operatorname{rank} B_1 = \operatorname{rank}[B_1, A_3]$；

3° 对于任意$\lambda \in \sigma(A_3)$，都有

$$\operatorname{rank}\begin{bmatrix} A_1 - \lambda I_{n_1} & A_3 \\ 0 & A_2 - \lambda I_{n_2} \\ C_1 & 0 \end{bmatrix} = n_1 + n_2.$$

5.3.3 动态补偿器的设计方法

动态补偿器的设计思想已在 5.3.1 小节中给出，这里给出具体步骤.

(1) 将控制输入向量 u 分为两部分

$$u = u_c + u_e, \tag{5-17}$$

按(A_1, B_1)能控性(或能稳性)，取 $m \times n_1$ 矩阵 K_1，使得 $A_1 + B_1K_1$ 为稳定矩阵.由条件(5-15)式确定矩阵 K_2，使得

$$A_3 = B_1K_2.$$

取控制规律

$$u_c = K_1x_1,$$
$$u_e = -K_2x_2.$$

由此得

$$u = K_1x_1 - K_2x_2. \tag{5-18}$$

(2) 由(C_1, A_1)的能观测性(或能检测性)和条件(5-16)式，设计状态观测器，取 $n_1 \times p$ 矩阵 G_1 和 $n_2 \times p$ 矩阵 G_2，使得

$$M = \begin{bmatrix} A_1 - GC_1 & A_3 \\ -G_2C_1 & A_2 \end{bmatrix}$$

为稳定矩阵.由此得状态观测器

$$\begin{bmatrix} \dot{x}_{1e} \\ \dot{x}_{2e} \end{bmatrix} = \begin{bmatrix} A_1 - G_1C_1 & A_3 \\ -G_2C_1 & A_2 \end{bmatrix}\begin{bmatrix} x_{1e} \\ x_{2e} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix}u + \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}y \tag{5-19}$$

和物理上易实现的反馈控制律

$$u = K_1x_{1e} - K_2x_{2e}, \tag{5-20}$$

其中 K_1 和 K_2 由(5-18)式确定.

(3) 设计带干扰补偿的动态补偿器.将控制规律(5-20)式代入系统(5-19)式，

然后和这个控制规律一起就可得到所要设计的动态补偿器：

$$\begin{cases}\dot{x}_{1e}=(A_1-G_1C_1+B_1K_1)x_{1e}+G_1y,\\ \dot{x}_{2e}=-G_2C_1x_{1e}+A_2x_{2e}+G_2y,\\ u=K_1x_{1e}-K_2x_{2e}.\end{cases}\tag{5-21}$$

例 1　已知二阶系统

$$\begin{cases}\dot{x}_1=-x_2,\\ \dot{x}_2=x_1+f+u,\\ y=x_1,\\ \dot{f}=0.\end{cases}\tag{5-22}$$

其中 x_1,x_2 是系统状态变量；u,y 分别为系统输入和输出；f 是外部干扰输入. 试对系统(5-22)式设计一个带干扰补偿的动态补偿器.

解　系统(5-22)式能写成如下向量形式：

$$\begin{cases}\dot{x}=Ax+A_3f+bu,\\ y=cx,\\ \dot{f}=0,\end{cases}\tag{5-23}$$

其中 $x=[x_1,x_2]^{\mathrm{T}},b=[0,1]^{\mathrm{T}},c=[1,0]$,

$$A=\begin{bmatrix}0&-1\\1&0\end{bmatrix},\quad A_3=\begin{bmatrix}0\\1\end{bmatrix}.$$

取 $K_1\in R^{1\times 2}$，使 $A+bK_1$ 是稳定的，例如可取

$$K_1=[0\quad -1].$$

取 K_2 使得 $A_3=bK_2$，得

$$K_2=1.$$

作系统

$$\begin{bmatrix}\dot{x}\\ \dot{f}\end{bmatrix}=\begin{bmatrix}A&A_3\\0&0\end{bmatrix}\begin{bmatrix}x\\ f\end{bmatrix}+\begin{bmatrix}0\\1\\0\end{bmatrix}u,$$

$$y=[1,0,0]$$

的状态观测器. 为此，取 G_1 和 G_2 使矩阵

$$M_1=\begin{bmatrix}A-G_1c&A_3\\-G_2c&0\end{bmatrix}$$

稳定，例如可取

$$G_1=\begin{bmatrix}3\\-2\end{bmatrix},\quad G_2=-1.$$

根据(5-21)式可得所求的动态补偿器为

$$\begin{cases}\dot{x}_{1e}=\begin{bmatrix}-3 & -1\\ 3 & -1\end{bmatrix}x_{1e}+\begin{bmatrix}3\\ -2\end{bmatrix}y,\\ \dot{x}_{2e}=[1,0]x_{1e}-y,\\ u=[0,-1]x_{1e}-x_{2e}.\end{cases} \tag{5-24}$$

5.4 内模原理

5.4.1 内模原理的初步引论

所谓内模原理,是为设计具有某种抗干扰能力的调节器所确定的一般原理.一个性能良好的调节器,它不仅在系统的标称参数处使闭环系统内部稳定和输出调节,而且应该在这些标称参数发生微小变化时,也能保持内部稳定和输出调节.具有这种性质的调节器称为"鲁棒调节器".内模原理就是设计这种鲁棒调节器所应遵循的一种原则.下面用一个简单的调节问题来说明内模原理的直观含义.

设一个单输入单输出的开环系统如图 5-1 所示,其传递函数为

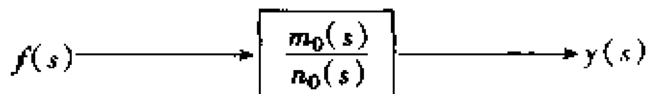


图 5-1

$$w_0(s)=\frac{m_0(s)}{n_0(s)},$$

其中 $m_0(s)$和 $n_0(s)$是 s 的多项式,$n_0(s)$首项系数为 1,$m_0(s)$与 $n_0(s)$互质,$\deg(m_0(s))=\alpha$,$\deg(n_0(s))=\beta$,$\alpha<\beta$,$\deg(\cdot)$表示多项式的次数,$f(s)$是系统的干扰输入.设 $f(t)$满足 k 阶微分方程

$$p(D)\,f(t)=0, \tag{5-25}$$

其中 $p(s)$是 s 的 k 次多项式,$D=\frac{\mathrm{d}}{\mathrm{d}t}$.设 $p(s)$的零点都在闭的右半复平面内.对方程(5-25)式作拉普拉斯变换,得到

$$f(s)=\frac{q(s)}{p(s)},$$

其中 $q(s)$是一个次数至多为 $k-1$ 的多项式,它由 $f(t)$及其导数的初值所确定,不同的初值决定不同的 $q(s)$.不失一般性,设 $q(s)$与 $p(s)$互质.

取系统的一个动态补偿器为

$$u(s)=-\frac{m_1(s)}{n_1(s)}y(s),$$

其中 $m_1(s)$与 $n_1(s)$都是 s 的多项式,$\deg(m_1(s))=\alpha_1$,$\deg(n_1(s))=\beta_1$,$\alpha_1\leqslant\beta_1$,$n_1(s)$的首项系数为 1,$m_1(s)$与 $n_1(s)$互质,则闭环系统如图 5-2 所示.

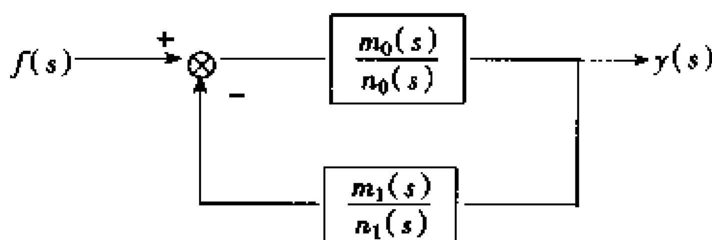


图 5-2

该闭环系统的传递函数是

$$W_c(s)=\frac{n_1(s)m_0(s)}{n_0(s)n_1(s)+m_0(s)m_1(s)}.$$

于是有

$$\begin{aligned}y(s)&=\frac{n_1(s)m_0(s)}{n_1(s)n_0(s)+m_1(s)m_0(s)}\cdot f(s)\\&=\frac{n_1(s)m_0(s)}{n_1(s)n_0(s)+m_1(s)m_0(s)}\frac{q(s)}{p(s)}.\end{aligned}\tag{5-26}$$

由于 $\alpha<\beta,\alpha_1\leqslant\beta_1$,故

$$\deg(n_1(s)n_0(s)+m_1(s)m_0(s))=\beta+\beta_1.$$

由此说明闭环系统是非退化的.

为了达到输出调节的目的,有理分式

$$\frac{n_1(s)m_0(s)}{n_1(s)n_0(s)+m_1(s)m_0(s)}\frac{q(s)}{p(s)}$$

的不稳定极点必须是可去极点.因此,必须有

$$p(s)\mid n_1(s)m_0(s),$$

其中 $A\mid B$ 表示 A 整除 B.

因为开环系统的标称参数可以发生微小变化,发生变化后的一般情况是 $m_0(s)$与 $p(s)$互质,于是必有

$$p(s)\mid n_1(s).\tag{5-27}$$

(5-27)式的控制理论意义是:一个"鲁棒控制器"的极点必须包含外部系统的不稳定极点.不太严格地说,极点代表着系统的动力学模型.因此(5-27)式也意味着"鲁棒控制器"的动力学模型编入了外部系统的不稳定的动力学模型.这就是内模原理的一般含义.

5.4.2 鲁棒调节器和内模原理

下面给鲁棒调节器和内模原理以明确的含义.

定义 2 给定系统(5-1)式.按一定次序将矩阵 A_1,B_1,C_1 的元素排成 $\mathbf{R}^v$ 空间的一个向量,$v=n_1^2+n_1m+n_1p$,那么这个向量称为系统(5-1)式的一个数据点,用 $P(A_1,B_1,C_1)$表示.

类似地定义数据点 $P(A_1,B_1,A_3)$等.

定义 3　设系统(5-4)式是系统(5-1)式的一个综合.如果有数据点 $P(A_1,B_1,C_1)$的某个邻域 U_p,使得当系统(5-1)式的参数在 U_p 内变化时,系统(5-4)式仍然是系统(5-1)式的综合,那么就称(5-4)式为系统(5-1)式在数据点 $P(A_1,B_1,C_1)$处的一个**结构稳定的综合**,也称为系统(5-1)式的一个鲁棒调节器.

定义 4　给定系统(5-1)式.如果存在 $q\times p$ 矩阵 Q,使得 D_1, C_1 和 D_2, C_2 满足 $D_1=QC_1$ 和 $D_2=QC_2$,则称 z 能从 y 读出.

定义 5　设给定系统(5-1)式和(5-4)式.如果 A_2 的最小多项式能整除 A_c 的 q 个不变因子,那么称 A_c 编入了 A_2 的一个内模.

例 2　设 A_2 的最小多项式为

$$f(\lambda)=\lambda^k+a_k\lambda^{k-1}+\cdots+a_2\lambda+a_1,$$

其对应的相伴标准形为

$$Q=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_k \end{bmatrix}.$$

若取

$$A_c=\mathrm{diag}\{\overbrace{Q,Q,\cdots,Q}^{q\text{个}}\},$$

则

$$A_c-\lambda I_l\triangleq\mathrm{diag}\{1,1,\cdots,1,\overbrace{f(\lambda),f(\lambda),\cdots,f(\lambda)}^{q\text{个}}\}.$$

其中 I_l 表示 l 阶单位矩阵.因此,$f(\lambda)$能整除 A_c 的 q 个不变因子,即 A_c 编入了 A_2 的一个内模.

下面是 A_c 编入 A_2 一个内模的判别准则.

引理 2　考虑系统(5-1)式和(5-4)式,A_c 编入 A_2 一个内模的充分必要条件是,对于任意的 $\lambda\in Z(A_2)$,都有

$$d(\ker(A_c-\lambda I_c)\cap\mathrm{Im}(A_c-\lambda I_l)^{k_\lambda-1})\geqslant q,$$

其中 $Z(A_2)$表示 A_2 的最小多项式的零点集,$d(\cdot)$为向量空间$(\cdot)$的维数,k_λ 为λ 的重数,q 为被调量 z 的维数,$\mathrm{Im}z$ 表示 z 的虚部.

定义 6　给定系统(5-1)式和(5-4)式,如果存在 $l\times l$ 非奇异矩阵 T,使得

$$T^{-1}A_cT=\begin{bmatrix} A_{c1} & A_{c3} \\ 0 & A_{c2} \end{bmatrix},\quad T^{-1}B_c=\begin{bmatrix} B_{c1} & B_{c3} \\ 0 & B_{c2} \end{bmatrix},$$

并且 A_{c2}编入了 A_2 的一个内模,其中,A_{c1} A_{c2} A_{c3}分别是 $l_1\times l_1$, $l_1\times l_2$ 和 $l_2\times l_2$ 矩阵,$l_1+l_2=l$, B_{c1}, B_{c2}, B_{c3}分别是 $l_1\times(p-q)$, $l_2\times q$ 和 $l_1\times q$ 矩阵,且如果对于每个 $\lambda\in\sigma(A_2)$,都有

$$\mathrm{rank}\begin{bmatrix} A_c-\lambda I_l \\ F_c \end{bmatrix}=l,$$

则称此内模关于 $\boldsymbol{u}$ 能观测;如果对于每个 $\lambda\in\sigma(\boldsymbol{A}_2)$,都有

$$\operatorname{rank}[\boldsymbol{A}_{c2}-\lambda\boldsymbol{I}_{l_2},\boldsymbol{B}_{c2}]=l_2,$$

则称此内模关于 z 能控.

根据上面这些概念,可得出下面两定理.

定理 3 设给定系统(5-1)式和(5-4)式.系统(5-4)式在数据点 $\boldsymbol{P}(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_3)$ 处是系统(5-1)式的结构稳定综合(或鲁棒调节器)的充分必要条件为

1° $(\boldsymbol{A}_1,\boldsymbol{B}_1)$能稳,且$(\boldsymbol{C}_1,\boldsymbol{A}_1)$能检测;

2° z 能从 $\boldsymbol{y}$ 读出;

3° 闭环系统内部稳定;

4° 系统(5-4)式编入了 $\boldsymbol{A}_2$ 的一个内模,并且该内模关于 $\boldsymbol{u}$ 能观测,关于 z 能控.

注:上述数据点 $\boldsymbol{P}(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_3)$的数据可以增加一些.例如将定理 2 中 $\boldsymbol{P}(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_3)$换为 $\boldsymbol{P}(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_3,\boldsymbol{C}_1)$后其结论亦成立.

定理 4 设给定系统(5-1)式.设 z 能从 $\boldsymbol{y}$ 读出,则在数据点 $\boldsymbol{P}(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_3)$处存在结构稳定的综合的充分必要条件为

1° $(\boldsymbol{A}_1,\boldsymbol{B}_1)$能稳,且$(\boldsymbol{C}_1,\boldsymbol{A}_1)$能检测;

2° 对于每个$\lambda\in\sigma(\boldsymbol{A}_2)$,都有

$$\operatorname{rank}\begin{bmatrix}\boldsymbol{A}_1-\lambda\boldsymbol{I}_1 & \boldsymbol{B}_1\\ \boldsymbol{D}_1 & \boldsymbol{0}\end{bmatrix}=n_1+q.$$

5.4.3 鲁棒调节器的设计

设定理 4 的条件满足.由于 z 能从 $\boldsymbol{y}$ 读出,不失一般性,可设 $\boldsymbol{C}_1=\begin{bmatrix}\boldsymbol{E}_1\\ \boldsymbol{D}_1\end{bmatrix}$,$\boldsymbol{C}_2=\begin{bmatrix}\boldsymbol{E}_2\\ \boldsymbol{D}_2\end{bmatrix}$,其中 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$ 分别为$(p-q)\times q$ 和$(p-q)\times n_2$ 矩阵.

鲁棒调节器的设计步骤如下.

(1) 求 $\boldsymbol{A}_2$ 的最小多项式 $f(\lambda)$.记为

$$f(\lambda)=\lambda^k+a_k\lambda^{k-1}+\cdots+a_2\lambda+a_1.$$

作相应于 $f(\lambda)$的相伴矩阵

$$\boldsymbol{Q}=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ -a_1 & -a_2 & -a_3 & \cdots & -a_k\end{bmatrix}.$$

(2) 设计伺服补偿器.取

$$\boldsymbol{A}_{c2}=\operatorname{diag}\{\overbrace{\boldsymbol{Q},\boldsymbol{Q},\cdots,\boldsymbol{Q}}^{q\text{个}}\},$$

$$B_{c2}=\mathrm{diag}\{\overbrace{e,e,\cdots,e}^{q个}\},\quad e=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}_{k\times 1}.$$

构造系统

$$\dot{x}_{c2}=A_{c2}x_{c2}+B_{c2}z,$$

上式即为**伺服补偿器**.

显然(A_{c2},B_{c2})能控.

(3) 设计镇定补偿器．考虑复合系统

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_{c2}\end{bmatrix}=\begin{bmatrix}A_1 & 0\\ B_{c2}D_1 & A_{c2}\end{bmatrix}\begin{bmatrix}x_1\\ x_{c2}\end{bmatrix}+\begin{bmatrix}B_1\\ 0\end{bmatrix}u,$$
$$\begin{bmatrix}y\\ x_{c2}\end{bmatrix}=\begin{bmatrix}C_1 & 0\\ 0 & I_{l_2}\end{bmatrix}\begin{bmatrix}x_1\\ x_{c2}\end{bmatrix},\tag{5-28}$$

其中 $l_2=kq$, u 是输入, $\begin{bmatrix}y\\ x_{c2}\end{bmatrix}$是输出.

由于(A_1,B_1)能稳,所以复合系统(5-28)式能稳的充分必要条件是,对于每一个 $\lambda\in\sigma(A_2)$,有

$$\mathrm{rank}\begin{bmatrix}A_1-\lambda I_{n_1} & 0 & B_1\\ B_{c2}D_1 & A_2-\lambda I_{l_2} & 0\end{bmatrix}=n_1+l_1,\tag{5-29}$$

因为

$$\mathrm{rank}\begin{bmatrix}A_1-\lambda I_{n_1} & 0 & B_1\\ B_{c2}D_1 & A_2-\lambda I_{l_2} & 0\end{bmatrix}$$
$$=\mathrm{rank}\begin{bmatrix}I_{n_1} & 0 & 0\\ 0 & B_{c2} & A_{c2}-\lambda I_{l_2}\end{bmatrix}\begin{bmatrix}A_1-\lambda I_{n_1} & B_1 & 0\\ D_1 & 0 & 0\\ 0 & 0 & I_{l_2}\end{bmatrix}.\tag{5-30}$$

由定理 3 的条件 2° 及(A_{c2},B_{c2})的能控性可知,

$$\mathrm{rank}\begin{bmatrix}I_{n_1} & 0 & 0\\ 0 & B_{c2} & A_{c2}-\lambda I_{l_2}\end{bmatrix}=n_1+l_2,$$

$$\mathrm{rank}\begin{bmatrix}A_1-\lambda I_{n_1} & B_1 & 0\\ D_1 & 0 & 0\\ 0 & 0 & I_q\end{bmatrix}=n_1+l_2+q.$$

由上式及(5-30)式可知,等式(5-29)式成立,即系统(5-28)式能稳和能检测.

这样就可以对复合系统(5-28)式设计一个使闭环系统为稳的镇定补偿器,并记为

$$
\dot{x}_{c1} = A_{c1} x_{c1} + [A_{c3} \quad \tilde{B}_c] \begin{bmatrix} x_{c2} \\ y \end{bmatrix},
\qquad (5\text{-}31)
$$

$$
u = F_{c1} x_{c1} + [F_{c2} \quad F] \begin{bmatrix} x_{c2} \\ y \end{bmatrix}.
$$

(4) 构成鲁棒调节器. 根据(1),(2)和(3)中所设定的数据,作动态补偿器:

$$
\begin{bmatrix} \dot{x}_{c1} \\ \dot{x}_{c2} \end{bmatrix} = \begin{bmatrix} A_{c1} & A_{c3} \\ 0 & A_{c2} \end{bmatrix} \begin{bmatrix} x_{c1} \\ x_{c2} \end{bmatrix} + \begin{bmatrix} B_{c1} & B_{c3} \\ 0 & B_{c2} \end{bmatrix} y, \qquad (5\text{-}32)
$$

$$
u = F_c x_c + F y.
$$

其中 $\tilde{B}_c = [B_{c1}, B_{c3}]$, $F_c = [F_{c1}, F_{c2}]$. 系统(5-32)式即为所求的鲁棒调节器.

例 3 设给定系统

$$
\begin{cases}
\dot{x}_1 = -x_2 + \varphi(t), \\
\dot{x}_2 = x_1 + u, \\
y = x_1, \\
\dot{\varphi}(t) = 0,
\end{cases}
\qquad (5\text{-}33)
$$

其中 u 是控制输入, y 是量测输出,也是被调节输出, $\varphi(t)$ 是外部阶跃干扰. 试对系统(5-33)式设计一个鲁棒调节器.

解 显然系统(5-30)式能控且能观测. 另有

$$
\operatorname{rank} \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = 3,
$$

因此,依定理 4,系统(5-33)式存在鲁棒调节器,其设计步骤如下.

(1) 计算外部干扰的最小多项式. 本题中

$$
f(s) = s.
$$

(2) 设计伺服补偿器. 因为被调节变量只有一个,而且 $\sigma(A_2) = \{0\}$,所以取伺服补偿器为

$$
\dot{x}_{c2}(t) = y(t).
$$

(3) 设计镇定补偿器. 作复合系统

$$
\begin{cases}
\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_{c2} \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_{c2} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u, \\
\begin{bmatrix} y \\ x_{c2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_{c2} \end{bmatrix}.
\end{cases}
\qquad (5\text{-}34)
$$

不难验证,复合系统(5-34)式能控且能观测. 因此(5-34)式能用动态补偿器镇定. 例如可用标点配置加状态观测器的方式镇定之.

$$\dot{x}_{c1}=\left(\begin{bmatrix}0&-1&0\\1&0&0\\1&0&0\end{bmatrix}-G\begin{bmatrix}1&0&0\\0&0&1\end{bmatrix}\right)x_{c1}+G\begin{bmatrix}y\\x_{c2}\end{bmatrix}+\begin{bmatrix}0\\1\end{bmatrix}u,$$

$$u=k^{\mathrm{T}}x_{c1}.$$

经计算,在本例中可取

$$G=\begin{bmatrix}3&0\\2&1\\0&0\end{bmatrix},\quad k^{\mathrm{T}}=[-2,-3,1].$$

(4) 综上所述,所求的一个鲁棒调节器可取为

$$\begin{cases}\begin{bmatrix}\dot{x}_{c1}\\\dot{x}_{c2}\end{bmatrix}=\begin{bmatrix}-3&-1&0&0\\-3&-3&2&1\\1&0&0&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}x_{c1}\\x_{c2}\end{bmatrix}+\begin{bmatrix}3\\2\\0\\1\end{bmatrix}y,\\u=[-2,-3,1]x_{c1}.\end{cases}\tag{5-35}$$

6　干扰解耦和无交互作用控制

6.1　干扰解耦问题的描述

考虑线性控制系统

$$\begin{cases}\dot{x}=Ax+Bu+Sq(t)\quad(t\geqslant 0),\\z=Dx,\end{cases}\tag{6-1}$$

其中 $x\in\mathbf{R}^n,u\in\mathbf{R}^m,z\in\mathbf{R}^p,q(t)\in\mathbf{R}^l,t\geqslant 0;A\in\mathbf{R}^{n\times n},B\in\mathbf{R}^{n\times m},D\in\mathbf{R}^{p\times n},S\in\mathbf{R}^{n\times l}$.

系统(6-1)式中的 $q(t)$表示作用于系统的干扰.假设干扰是不能直接量测的.

这里要解决的问题是:求(如果可能)线性状态反馈 $u=Fx$,使得干扰 $q(t)$对闭环系统

$$\begin{cases}\dot{x}=(A+BF)x+Sq(t),\\z=Dx\end{cases}\tag{6-2}$$

的输出 $z(t)$没有影响.这里"$q(t)$对 $z(t)$的影响"与 $q(t)$的变化范围有关.

假定 Q 是定义在$[0,+\infty)$上取值于 $\mathbf{R}^l$ 中连续函数的全体组成的集合.如果对于每个初态 $x(0)$,系统(6-2)式的输出 $z(t)$对于任意 $q(\cdot)\in Q$ 都相同,则称系统(6-2)式的输出 $z(t)$是干扰解耦的.简称系统(6-2)式是干扰解耦的.于是,干扰解耦意味着:对于所有的 $q(\cdot)\in Q$,皆成立

$$z(t)=D\int_0^t\exp((t-\tau)(A+BF))Sq(\tau)\mathrm{d}\tau=0(t\geqslant 0).\tag{6-3}$$

记 $\mathscr{K}=\ker \boldsymbol{D}$，$\mathscr{S}=\operatorname{Im}\boldsymbol{S}$，则从(6-3)式可证得出如下引理.

引理 1 系统(6-2)式干扰解耦的充分必要条件为

$$\langle \boldsymbol{A}+\boldsymbol{BF}\mid\mathscr{S}\rangle\subset\mathscr{K}. \tag{6-4}$$

其中$\langle \boldsymbol{A}+\boldsymbol{BF}\mid\mathscr{S}\rangle$的定义参见(2-12)式.

于是，用状态反馈使干扰解耦的问题的提法如下.

干扰解耦问题(DDP)：设 $\boldsymbol{A}:\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{B}:\mathbf{R}^m\to\mathbf{R}^n$，子空间 $\mathscr{S}\subset\mathbf{R}^n$ 和子空间 $\mathscr{K}\subset\mathbf{R}^n$. 求(如果可能)$\boldsymbol{F}:\mathbf{R}^n\to\mathbf{R}^m$，使得

$$\langle \boldsymbol{A}+\boldsymbol{BF}\mid\mathscr{S}\rangle\subset\mathscr{K}. \tag{6-5}$$

6.2 $(\boldsymbol{A},\boldsymbol{B})$不变子空间

6.2.1 $(\boldsymbol{A},\boldsymbol{B})$不变子空间的概念

$(\boldsymbol{A},\boldsymbol{B})$不变子空间的概念在干扰解耦等问题中起着关键性的作用.

定义 1 设 $\boldsymbol{A}:\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{B}:\mathbf{R}^m\to\mathbf{R}^n$. 如果对于子空间 $\mathscr{Y}\subset\mathbf{R}^n$，存在线性映射 $\boldsymbol{F}:\mathbf{R}^n\to\mathbf{R}^m$，使得

$$(\boldsymbol{A}+\boldsymbol{BF})\mathscr{Y}\subset\mathscr{Y},$$

则称 $\mathscr{Y}$ 是$(\boldsymbol{A},\boldsymbol{B})$不变子空间.

设 $\mathscr{K}\subset\mathbf{R}^n$ 为子空间. 用 $\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$表示包含于 $\mathscr{K}$ 中的所有$(\boldsymbol{A},\boldsymbol{B})$不变子空间组成的族，即

$$\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})=\{\mathscr{Y}\mid\mathscr{Y}\text{为}(\boldsymbol{A},\boldsymbol{B})\text{不变子空间},\mathscr{Y}\subset\mathscr{K}\}. \tag{6-6}$$

6.2.2 最大$(\boldsymbol{A},\boldsymbol{B})$不变子空间

$\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$是非空的，因为零空间显然属于它. $\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$关于子空间的加法是封闭的，即

$$\mathscr{Y}_1+\mathscr{Y}_2\in\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})\quad(\forall\,\mathscr{Y}_1,\mathscr{Y}_2\in\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})),$$

因此，$\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$关于子空间的加法的包含关系是一个上半格. 由此容易知道，存在唯一的 $\mathscr{Y}^*\in\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$，具有下列性质：

$$\mathscr{Y}\subset\mathscr{Y}^*\quad(\forall\,\mathscr{Y}\in\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})). \tag{6-7}$$

定义 2 称满足(6-7)式的 Y^* 为 $\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$的最大元，或 $\mathscr{K}$ 中的**最大$(\boldsymbol{A},\boldsymbol{B})$不变子空间.**

记

$$\mathscr{Y}^*=\sup\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K}). \tag{6-8}$$

6.3 干扰解耦问题可解性条件

定理 1 干扰解耦问题(DDP)可解的充分必要条件为

$$\mathscr{Y}^*\supset\mathscr{S}, \tag{6-9}$$

其中 $\mathscr{Y}^*=\sup\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$.

6.4　最大$(\boldsymbol{A},\boldsymbol{B})$不变子空间的计算

1. 几何计算法

设 $\boldsymbol{A}:\mathbf{R}^n\to\mathbf{R}^n$，$\boldsymbol{B}:\mathbf{R}^m\to\mathbf{R}^n$，子空间 $\mathscr{K}\subset\mathbf{R}^n$. 定义序列 $\mathscr{Y}^\mu$ 如下：

$$\mathscr{Y}^0=\mathscr{K},$$

$$\mathscr{Y}^\mu=\mathscr{K}\cap\boldsymbol{A}^{-1}(\mathscr{B}+\mathscr{Y}^{\mu-1})\quad(\mu=1,2,\cdots),$$

则必在第 k 步（$k\leqslant d(\mathscr{K})$，$d(\mathscr{K})$为子空间 $\mathscr{K}$ 的维数）时，有

$$\mathscr{Y}^{k+1}=\mathscr{Y}^k,$$

到此停止计算，并且有

$$\mathscr{Y}^k=\sup\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K}).$$

2. 代数计算法

先引入矩阵方程最大解的概念：给定矩阵 $\boldsymbol{M},\boldsymbol{X},\boldsymbol{Y}$，所谓方程 $\boldsymbol{MX}=0$（或 $\boldsymbol{YM}=0$）的最大解，是指具有最大秩的一个解 $\boldsymbol{X}$（或 $\boldsymbol{Y}$），其不为零的列（或行）是线性独立的.

代数计算方法如下.

取 $\mathscr{K}=\ker\boldsymbol{D}$. 定义矩阵序列 $\boldsymbol{V}_\mu,\boldsymbol{W}_\mu$ 如下：

(1) $\boldsymbol{V}_0$ 为 $\boldsymbol{DV}_0=0$ 的最大解；

(2) $\boldsymbol{W}_\mu$ 是 $\boldsymbol{W}_\mu[\boldsymbol{B},\boldsymbol{V}_{\mu-1}]$的最大解，$\mu=1,2,\cdots$；

(3) $\boldsymbol{V}_\mu$ 是方程$\begin{bmatrix}\boldsymbol{D}\\ \boldsymbol{W}_\mu\boldsymbol{A}\end{bmatrix}\boldsymbol{V}_\mu=0$，的最大解，$\mu=1,2,\cdots$；

(4) 检验 $\operatorname{rank}[\boldsymbol{V}_{\mu-1},\boldsymbol{V}_\mu]=\operatorname{rank}\boldsymbol{V}_{\mu-1}$是否成立，不成立转到第(2)步；成立，则停止计算.

第(2)，(3)和(4)步的循环计算必在第 k 步（$k\leqslant d(\mathscr{K})$）时停止. 此时有

$$\operatorname{rank}[\boldsymbol{V}_k,\boldsymbol{V}_{k+1}]=\operatorname{rank}\boldsymbol{V}_k,$$

和

$$\operatorname{Im}\boldsymbol{V}_k=\sup\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K}).$$

6.5　干扰解耦问题的求解

设 $\mathscr{S}\subset\mathscr{Y}^*=\sup\mathscr{T}(\boldsymbol{A},\boldsymbol{B};\mathscr{K})$成立. 下面给出求 $\boldsymbol{F}\in\mathbf{R}^{n\times m}$，使得$\langle\boldsymbol{A}+\boldsymbol{BF}|\mathscr{S}\rangle\subset\langle\boldsymbol{A}+\boldsymbol{BF}|\mathscr{Y}^*\rangle\subset\mathscr{K}$的方法.

因为 $\mathscr{Y}^*$ 是$(\boldsymbol{A},\boldsymbol{B})$不变子空间，所以

$$\boldsymbol{A}\mathscr{Y}^*\subset\mathscr{Y}^*+\mathscr{B}.\tag{6-10}$$

设 $\mathscr{Y}^*$ 的一组基为$\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k\}$. 由(6-10)式可知，存在 $\boldsymbol{w}_i\in\mathscr{Y}^*$ 和 $\boldsymbol{u}_i\in\mathbf{R}^m$，$i=1,2,\cdots,k$，使得

$$\boldsymbol{Av}_i=\boldsymbol{w}_i+\boldsymbol{Bu}_i\quad(i=1,2,\cdots,k).\tag{6-11}$$

定义 $\boldsymbol{F}_0:\mathscr{Y}^*\to\mathbf{R}^m$，

$$F_0 v_i = -u_i \quad (i=1,2,\cdots,k). \tag{6-12}$$

设 F 是 F_0 在 $\mathbf{R}^n$ 上的扩张.

由(6-11)式和(6-12)式及 F 是 F_0 在 $\mathbf{R}^n$ 上的扩张,可得

$$\begin{aligned}(A+BF)v_i &= Av_i + BFv_i \\ &= w_i + Bu_i + BF_0 v_i \\ &= w_i + Bu_i + B(-u_i) \\ &= w_i. \quad (i=1,2,\cdots,k),\end{aligned}$$

由此得到

$$\begin{aligned}(A+BF)\mathscr{V}^* &= (A+BF)\mathrm{span}\{v_1,v_2,\cdots,v_k\} \\ &= \mathrm{span}\{w_1,w_2,\cdots,w_k\} \subset \mathscr{V}^*.\end{aligned}$$

因此,F 即为所求.

例 1 考虑系统(6-1)式,其中

$$A=\begin{bmatrix}0&1&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&0\end{bmatrix},\ B=\begin{bmatrix}0&0\\0&0\\1&0\\0&0\\0&1\end{bmatrix},$$

$$D=\begin{bmatrix}1&0&0&-1&0\\1&-1&0&0&0\\0&0&0&1&-1\end{bmatrix},\quad S=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}.$$

试分析该系统能否干扰解耦?如果能解耦,试求解耦反馈矩阵 F.

解 先根据 6.4 节中的几何计算法,求

$$\mathscr{V}^* = \sup\mathscr{I}(A,B;\ker D).$$

$$\ker D = \mathrm{span}\left\{\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}\right\},\quad \mathscr{B} = \mathrm{span}\left\{\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix},\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\},$$

$\mathscr{V}_0 = \ker D$,

$$\begin{aligned}\mathscr{V}_1 &= \ker D \cap A^{-1}(\mathscr{B}+\mathscr{V}_0) \\ &= \ker D \cap A^{-1}\mathrm{span}\left\{\begin{bmatrix}1\\1\\0\\0\\0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\}\end{aligned}$$

$$= \operatorname{span}\left\{\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}\right\}\cap\operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\1\\0\end{bmatrix}\begin{bmatrix}0\\1\\1\\1\\1\end{bmatrix}\right\}$$

$$= \operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}\right\},$$

$$\mathscr{Y}_2 = \ker \boldsymbol{D}\cap \boldsymbol{A}^{-1}(\mathscr{B}+\mathscr{Y}_1)$$

$$= \operatorname{span}\left\{\begin{bmatrix}1\\1\\0\\1\\1\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}\right\}\cap\operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\1\\0\end{bmatrix},\begin{bmatrix}0\\1\\1\\1\\1\end{bmatrix}\right\}$$

$$= \operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}\right\}.$$

因此,$\mathscr{Y}_2 = \mathscr{Y}_3$,所以

$$\mathscr{Y}^* = \operatorname{span}\left\{\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}\right\} = \operatorname{Im}\boldsymbol{S}.$$

由此可知,系统能解耦.

设 $\boldsymbol{q}_1 = \boldsymbol{s}$ 为 $\mathscr{Y}^*$ 的基,将 $\boldsymbol{q}$ 在 $\mathbf{R}^5$ 上扩张得 $\mathbf{R}^5$ 的一组基:$\boldsymbol{q}_1,\boldsymbol{q}_2,\boldsymbol{q}_3,\boldsymbol{q}_4,\boldsymbol{q}_5$,这里

$$\boldsymbol{q}_2=\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix},\quad \boldsymbol{q}_3=\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix},\quad \boldsymbol{q}_4=\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix},\quad \boldsymbol{q}_5=\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}.$$

通过计算得

$$\boldsymbol{A}\boldsymbol{q}_1 = \boldsymbol{q}_1 - \boldsymbol{B}\begin{bmatrix}1\\1\end{bmatrix}.$$

定义 $\boldsymbol{F}:\mathbf{R}^5\to\mathbf{R}^2$,

$$Fq_1=\begin{bmatrix}1\\1\end{bmatrix},\quad Fq_i=0\quad(i=2,3,4,5).$$

由此得

$$F=\begin{bmatrix}1&0&0&0&0\\1&0&0&0&0\end{bmatrix}.$$

6.6 带有稳定性的干扰解耦

6.6.1 问题的提法

设 $A:\mathbf{R}^n\to\mathbf{R}^n$, $B:\mathbf{R}^m\to\mathbf{R}^n$, $\mathscr{S}\subset\mathbf{R}^n$, $\mathscr{K}\subset\mathbf{R}^n$.

带有稳定性的干扰解耦(DDPS)是指:求 $F:\mathbf{R}^n\to\mathbf{R}^m$ 使得

$$\langle A+BF|\mathscr{S}\rangle\subset\mathscr{K},\quad \sigma(A+BF)\subset C^-. \tag{6-13}$$

6.6.2 能控性子空间

定义 3 设子空间 $\mathscr{R}\subset\mathbf{R}^n$. 如果存在 $F:\mathbf{R}^n\to\mathbf{R}^m$ 和 $G:\mathbf{R}^m\to\mathbf{R}^m$ 使得

$$\mathscr{R}=\langle A+BF|\operatorname{Im}(BG)\rangle, \tag{6-14}$$

则称 $\mathscr{R}$ 为(A,B)的**能控性子空间**.

用$G(A,B;\mathscr{K})$表示包含于 $\mathscr{K}$ 的(A,B)能控性子空间组成的族. $G(A,B;\mathscr{K})$存在一个最大的(A,B)能控性子空间 $\mathscr{R}^*$,记为

$$\mathscr{R}^*=\sup G(A,B;\mathscr{K}). \tag{6-15}$$

关于最大(A,B)不变子空间 $\mathscr{V}^*$ 与最大(A,B)能控性子空间 $\mathscr{R}^*$之间有下面的关系.

引理 2 设 $\mathscr{V}^*=\sup\mathscr{T}(A,B;\mathscr{K})$, $\mathscr{R}^*=G(A,B;\mathscr{K})$. 设 $F:\mathbf{R}^n\to\mathbf{R}^m$,使得$(A+BF)\mathscr{V}^*\subset\mathscr{V}^*$,则有

$$\mathscr{R}^*=\langle A+BF|\mathscr{B}\cap\mathscr{V}^*\rangle,\quad \mathscr{B}=\operatorname{Im}B. \tag{6-16}$$

6.6.3 DDPS 可解性条件

设

$$\mathscr{V}^*=\sup\mathscr{T}(A,B;\mathscr{K}),$$
$$\mathscr{R}^*=\sup G(A,B;\mathscr{K}).$$

取 $F_0:\mathbf{R}^n\to\mathbf{R}^m$,使得

$$(A+BF_0)\mathscr{V}^*\subset\mathscr{V}^*. \tag{6-17}$$

记 $A_0=A+BF_0$. 设 $P:\mathbf{R}^n\to\dfrac{\mathbf{R}^n}{\mathscr{R}^*}$是标准投影. 设 $\bar{A}_0$ 是在$\dfrac{\mathbf{R}^n}{\mathscr{R}^*}$中由 A_0 产生的诱导映射.因为

$$\bar{A}_0\left(\frac{\mathscr{V}^*}{\mathscr{R}^*}\right)=\bar{A}_0P\mathscr{V}^*=PA_0\mathscr{V}^*\subset P\mathscr{V}^*=\frac{\mathscr{V}^*}{\mathscr{R}^*},$$

所以$\dfrac{\mathscr{V}^*}{\mathscr{R}^*}$是 $\bar{A}_0$ 不变的.

另外易证，$\bar{A}_0$ 在 $\frac{\mathscr{V}^*}{\mathscr{R}^*}$ 上的限制同满足(6-17)式的 F_0 的选择无关. 设 $\alpha(\lambda)$ 是 $A_0 \Big| \left(\frac{\mathscr{V}^*}{\mathscr{R}^*}\right)$ 的最小多项式. 设

$$\alpha(\lambda) = \alpha_g(\lambda)\ \alpha_b(\lambda),$$

其中 $\alpha_g(\lambda)$($\alpha_b(\lambda)$)的零点属于 C^-(C^+). 记

$$\bar{\mathscr{X}}_g^* = \frac{\mathscr{V}^*}{\mathscr{R}^*} \cap \ker \alpha_g(\bar{A}_0).$$

定义

$$\mathscr{V}_g^* = P^{-1}\bar{\mathscr{X}}_g^*. \tag{6-18}$$

定理 2 DDPS 可解的充分必要条件为

$$\mathscr{S} \subset \mathscr{V}_g^*. \tag{6-19}$$

关于最大(A,B)能控性子空间 $\mathscr{R}^*$ 的算法和反馈矩阵 F 的求法见文献[1].

6.7 无交互作用控制

6.7.1 问题提法

考虑线性控制系统

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx. \end{cases} \tag{6-20}$$

其中 $x \in \mathbf{R}^n$, $u, y \in \mathbf{R}^m$; $A \in \mathbf{R}^{n \times n}$, $B \in \mathbf{R}^{n \times m}$, $C \in \mathbf{R}^{m \times n}$.

取反馈律

$$u = Kx + Lv,$$

其中 $K \in \mathbf{R}^{m \times n}$, $L \in \mathbf{R}^{m \times m}$, 代入(6-20)式，得到

$$\dot{x} = (A + BK)x + BLv, \tag{6-21}$$

其中 v 是新的输入.

无交互作用控制问题：对系统(6-20)式求 $K \in \mathbf{R}^{m \times n}$ 和 $L \in \mathbf{R}^{m \times n}$(如果存在)，使得系统(6-21)式的传递函数阵 $G_{K,L}(s)$ 为非奇异对角有理分式矩阵.

如果这样的矩阵 K, L 存在，则称系统(6-20)式**无交互作用控制问题可解**.

无交互作用控制的含义是，每个控制分别只对一个输出变量有影响，因此，可用解单输入单输出系统的方法对系统进行设计.

6.7.2 无交互作用问题可解的条件

设 $G(s) = [g_{ij}(s)]_{m \times m}$ 为系统(6-20)式的传递函数矩阵. 下面定义两组特征量.

(1) 设

$k_{ij} = g_{ij}(s)$的分母多项式次数 $-$ $g_{ij}(s)$的

分子多项式次数 $(i,j=1,2,\cdots,m)$.

定义第一组特征量 d_i 为

$$d_i=\min\{k_{i1},k_{i2},\cdots,k_{im}\}\quad(i=1,2,\cdots,m).\tag{6-22}$$

(2) 设

$$g_i(s)=[g_{i1}(s),g_{i2}(s),\cdots,g_{im}(s)],$$

定义第二组特征量 E_i 为

$$E_i=\lim_{s\to\infty}s^{d_i}g_i(s)\quad(i=1,2,\cdots,m).\tag{6-23}$$

显然 $E_i(i=1,2,\cdots,m)$是 $1\times m$ 常数矩阵.

定理 3 系统(6-20)式的无交互作用问题可解的充分必要条件为

$$\boldsymbol{E}=\begin{bmatrix}\boldsymbol{E}_1\\ \boldsymbol{E}_2\\ \vdots\\ \boldsymbol{E}_m\end{bmatrix}\tag{6-24}$$

为非奇异矩阵,其中 $\boldsymbol{E}_i,i=1,2,\cdots,m$,由(6-23)式定义.

6.7.3 无交互作用控制问题的求解

设(6-24)式中 $\boldsymbol{E}$ 非奇异,反馈矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{L}$ 的求法如下.

(1) 计算由(6-22)式定义的$\{d_1,d_2,\cdots,d_m\}$,并按公式 $\boldsymbol{E}_i=\boldsymbol{C}_i\boldsymbol{A}^{d_i-1}\boldsymbol{B},i=1,2,\cdots,m$,计算$\{\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_m\}$,其中 $\boldsymbol{C}_i^{\mathrm{T}}$ 为 $\boldsymbol{C}$ 的第 i 行.

如果矩阵

$$\boldsymbol{E}=\begin{bmatrix}\boldsymbol{E}_1\\ \boldsymbol{E}_2\\ \vdots\\ \boldsymbol{E}_m\end{bmatrix}$$

奇异,则干扰解耦问题无解.如果 $\boldsymbol{E}$ 非奇异,则无交互作用控制问题有解,继续计算.

(2) 取

$$\boldsymbol{K}=-\boldsymbol{E}^{-1}\boldsymbol{F},\quad \boldsymbol{L}=-\boldsymbol{E}^{-1},\tag{6-25}$$

其中

$$\boldsymbol{F}=\begin{bmatrix}\boldsymbol{C}_1 & \boldsymbol{A}^{\alpha_1}\\ \boldsymbol{C}_2 & \boldsymbol{A}^{\alpha_2}\\ & \vdots\\ \boldsymbol{C}_m & \boldsymbol{A}^{\alpha_m}\end{bmatrix}.\tag{6-26}$$

(3) 写出解耦系统:

$$\begin{cases}\dot{\boldsymbol{x}}=(\boldsymbol{A}+\boldsymbol{BK})\boldsymbol{x}+\boldsymbol{BLv},\\ \boldsymbol{y}=\boldsymbol{Cx}.\end{cases}\tag{6-27}$$

例 2 判断系统

$$\dot{x}=\begin{bmatrix}3 & 1 & 0\\ 0 & 0 & -1\\ 0 & 1 & -1\end{bmatrix}x+\begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}u,$$

$$y=\begin{bmatrix}2 & -1 & 1\\ 0 & 2 & 1\end{bmatrix}x$$

无交互作用控制问题是否可解.

解　系统(6-27)式的传递函数阵为

$$G(s)=\begin{bmatrix}2 & -1 & 1\\ 0 & 2 & 1\end{bmatrix}\begin{bmatrix}s-3 & -1 & 0\\ 0 & s & 1\\ 0 & -1 & s+1\end{bmatrix}^{-1}\begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}$$

$$=\frac{1}{(s-3)(s^2+s+1)}\begin{bmatrix}-s^2+5s+2 & s^2-2s-5\\ 2s^2-3s-9 & s^2-5s+6\end{bmatrix}. \tag{6-28}$$

由(6-28)式可知,

$$d_1=d_2=1.$$

由此,得

$$E=CB=\begin{bmatrix}-1 & 1\\ 2 & 1\end{bmatrix}$$

非奇异,因此,无交互作用控制问题可解.

注:对于更一般的状态反馈无交互作用控制问题和用动态补偿器的无交互作用控制问题见文献[1].

参 考 文 献

1　旺纳姆 WM 著.线性多变量控制:一种几何方法.姚景尹,王恩平译.北京:科学出版社,1984.

2　王恩平,秦化淑,王世林.线性控制系统引论.广州:广东科技出版社,1991.

·经济数学卷·

第13篇

最优控制理论

编　者　秦化淑　王朝珠
审校者　王　翼

目　录

引　言

最优控制理论是现代控制理论中最早发展起来的分支之一.对于一个给定的受控系统,常常要求找到这样的控制函数,使得在它的作用下,系统从一个状态转移到为设计者希望的另一个状态,且使得系统的某种性能品质尽可能好.通常称这种问题为最优控制问题.**最优控制理论**是讨论最优控制(函数)应满足的必要条件,最优控制(函数)的存在和唯一性,以及求解最优控制问题等等的方法和理论.虽在经典控制理论中有过以系统的响应面积最小为指标的控制原理和方法,也有过以过渡时间最短为指标的"布绍原理",但真正形成控制系统以性能最优为目标的理论——最优控制理论,却是在20世纪50年代末.其主要标志是前苏联数学家庞特里亚金(L.C.Pontryagin)等人提出的"极大值原理".当受控对象的动力学模型由常微分方程描述时,有集中参数系统的最优控制理论;当受控对象的动力学模型由偏微分(积分)方程或随机微分方程描述时,则有分布参数系统或随机系统的最优控制理论.本篇主要介绍连续时间集中参数系统最优控制理论的若干问题.

1　最优控制问题

1.1　最优控制问题的几个实例

1. 升降机的快速下降问题

它是在工业和生活中常见的问题,如矿井中的提升机的升降,高层建筑中的电梯的升降等都可归为这类问题.

设有一升降机,升降机示意图如图1-1所示.记升降机为W,其质量为 m.

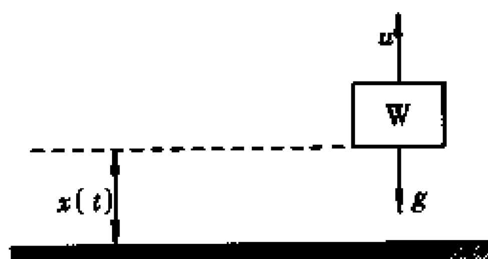


图 1-1

升降机W一方面受重力作用,另一方面受控制力作用.记重力为常数 g,而控制力为 u, u 通常为时间 t 的函数.实际中 u 是有限制的,即 $|u| \leqslant M$. M 为正常数.

为了保证控制力 u 能操纵升降机,显然应有 $M > g$.

设 $x(t)$ 为 t 时刻升降机距地面的高度,记初始时刻 $t_0 = 0$ 时,升降机 W 距地面高度为 $x(0) = x_1^0$,而垂直运动的速度为 $\dot{x}(0) = x_2^0$,这里 x_1^0, x_2^0 均为给定常数.

所谓升降机的快速下降问题就是如何选择控制力 $u(t)$,使得升降机最快到达地面,且到达地面时升降机的运动速度为零.记

$$x_1(t) = x(t), \quad x_2(t) = \dot{x}_1(t) = \dot{x}(t).$$

利用牛顿第二定律易得升降机的状态方程为

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = \frac{1}{m}(g - u(t)). \tag{1-1}$$

其初始条件为

$$x_1(0) = x_1^0, \quad x_2(0) = x_2^0. \tag{1-2}$$

如果记升降机到达地面的时刻为 t_f,通称终端时刻.据上面要求易知,(1-1)式的终端条件为

$$x_1(t_f) = 0, \quad x_2(t_f) = 0. \tag{1-3}$$

因此,升降机的快速下降问题就是选择一个满足

$$|u| \leqslant M \quad (M > g) \tag{1-4}$$

的控制力 $u(t)$,把升降机 W 由初态(1-2)式转移到终端状态(1-3)式,且使性能指标

$$J(u) = \int_0^{t_f} \mathrm{d}t = t_f \tag{1-5}$$

达到最小.

2. 生产计划问题

设 $t_f > 0$ 是生产计划结束的终端时刻.令 $x(t)$ 表示 t 时刻($0 \leqslant t \leqslant t_f$)商品库存量;$r(t) \geqslant 0$ 表示 t 时刻对商品的需求率,且为已知函数;$u(t)$ 表示 t 时刻的商品生产率,它将由计划人员来选取,是控制变量.如果对商品的上述要求能全部满足,则库存量 x 应满足

$$\dot{x} = -r(t) + u, \quad x(0) = x_0, \tag{1-6}$$

其中 $x(0)$ 是初始时刻 $t_0 = 0$ 的库存量.

由于 x 是库存量,u 是生产率,因此,必有

$$x(t) \geqslant 0 \qquad (0 \leqslant t \leqslant t_f), \tag{1-7}$$

$$0 \leqslant u(t) \leqslant M \qquad (0 \leqslant t \leqslant t_f), \tag{1-8}$$

其中 M 表示最高生产率.显然,为了使生产有序进行,M 应满足

$$M > r(t) \qquad (0 \leqslant t \leqslant t_f). \tag{1-9}$$

设 $b > 0$ 是单位时间内储存单位商品所需费用,而单位时间生产成本为生产率 $u(t)$ 的一个已知函数 $h(t, u(t))$.因此,该系统 t 时刻单位时间成本为

$$l(t, x(t), u(t)) = h(t, u(t)) + bx(t),$$

而时间间隔$[0, t_f]$内的总成本为

$$J(u) = \int_0^{t_f} l(t, x(t), u(t))\mathrm{d}t. \tag{1-10}$$

所谓生产计划问题,就是寻找一个最优生产率使总成本达到最小.

3. 防天拦截问题

现今国防上不但有防空问题,而且有防天问题,即防御洲际导弹和航天武器的问题.假设用拦截器L拦击来袭目标M.在某惯性坐标系中,记拦截器L和目标M的质心位置矢量分别为 $\boldsymbol{x}_L, \boldsymbol{x}_M$,如图 1-2 所示.

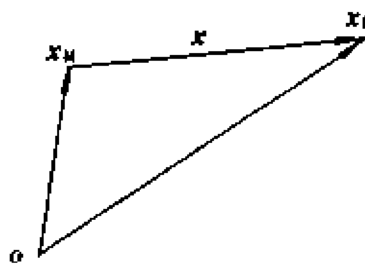


图 1-2

令相对位置矢量为

$$\boldsymbol{x} = \boldsymbol{x}_L - \boldsymbol{x}_M,$$

则相对速度矢量为

$$\boldsymbol{v} = \dot{\boldsymbol{x}}_L - \dot{\boldsymbol{x}}_M.$$

设 $\boldsymbol{a}(t)$ 是相对固有(除机动外的)加速度矢量,且简化为时间 t 的已知函数.若记拦截器L的质量为 $m(t)$,推力大小为 $f(t)$,推力方向为 $\boldsymbol{u}$,有效喷气速度为常数 c,则拦截器L关于目标M的相对运动方程为

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{v}, \\ \dot{\boldsymbol{v}} = \boldsymbol{a}(t) + \dfrac{f(t)}{m(t)}\boldsymbol{u}, \\ \dot{m}(t) = -\dfrac{f(t)}{c}. \end{cases} \tag{1-11}$$

而初始时刻 t_0 的状态为

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0, \quad \boldsymbol{v}(t_0) = \boldsymbol{v}_0, \quad m(t_0) = m_0. \tag{1-12}$$

考虑实际工程情况,则控制量 $f(t)$ 和 $\boldsymbol{u}(t)$ 应满足如下约束:

$$\begin{aligned} &0 \leqslant f(t) \leqslant \max f(t) \overset{\text{def}}{=\!=} F, \\ &\|\boldsymbol{u}\|_2^2 = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} = 1. \end{aligned} \tag{1-13}$$

由于拦截器L的质量不能小于所有燃料消耗完后的有效载荷 m_e,若记 t_f 为拦截过程结束(终端)时刻,则有

$$m(t_f) \geqslant m_e. \tag{1-14}$$

根据拦截的要求,其终端状态为

$$\boldsymbol{x}(t_f) = \boldsymbol{0}, \quad \boldsymbol{v}(t_f)\text{ 任意}. \tag{1-15}$$

为了使整个拦截过程时间尽量短,燃料尽量省,取性能指标为

$$J(f,u) = \int_{t_0}^{t_f} (c_1 + f(t))\mathrm{d}t, \tag{1-16}$$

其中 c_1 是常数，它是对时间的加权因子.

所谓最优拦截系指选择控制 (f,u)，使系统(1-11)式从初始状态(1-12)式出发，在终端时刻 t_f 时满足状态的终端要求(1-14)式、(1-15)式，且使性能指标(1-16)式达到最小.

1.2 最优控制问题

1.2.1 控制系统

控制系统通常是指描述被控对象动态行为的动力学模型. 这里指的是由常微分方程描述的动力学模型，即所论的控制系统由如下矢量微分方程描述：

$$\dot{x} = f(t,x,u), \quad x(t_0) = x_0. \tag{1-17}$$

其中 $x \in \mathbf{R}^n$ 是状态；$u \in \mathbf{R}^m$ 是控制；$x_0 \in \mathbf{R}^n$ 是初始状态，$f(\cdots)$ 是定义在 $[0, +\infty) \times \mathbf{R}^n \times \mathbf{R}^m$ 中某区域内的连续矢值函数，且将每个分段连续控制函数 $u(t)$ 代入后，初值问题(1-17)式存在唯一解.

1.2.2 约束条件

加在控制系统上的约束条件系指加在状态、控制变量取值上的限制，又称为**控制约束**. 通常控制变量都是一些能改变控制系统的动态行为，且为取值受限制的物理量，它可通过能力有限的实际物理装置来形成，因此，控制变量的取值应有限制，一般记为

$$u \in U \subseteq \mathbf{R}^m \qquad (U \text{ 为 } \mathbf{R}^m \text{ 中的子集，可闭、可开}). \tag{1-18}$$

加在状态变量取值上的限制，除少数情况外(如 1.1 节的 1. 中要求状态 $x(t) \geqslant 0$)，通常是指加在终端时刻状态的约束，它是由控制的目标所确定的，即要求系统终端状态 $x(t_f)$(t_f 是终端时刻)属于 $\mathbf{R}^n$ 中的某个子集 S. S 通常称为**目标集**. 一般有

$$S \xlongequal{\text{def}} \{x \mid g(x,t_f) = 0, \quad g \in \mathbf{R}^p, p < n\}. \tag{1-19}$$

一个分段连续矢值函数 $u(t)$，当它在(1-18)式的 U 中取值，并使初值问题(1-17)式的解 $x(t)$ 在某个区间 $[t_0,t_f]$ 上，$t_f > t_0$，存在，且唯一，而在终端时刻 t_f 的状态 $x(t_f)$ 满足(1-19)式时，则称 $u(t)$ 为**容许控制**. 容许控制的全体所组成的集合称为容许控制函数集，通常记为 $\mathscr{U}$.

1.2.3 性能指标

判别控制系统性能品质优劣的标准称为**性能指标**，通常记为 $J(u(\cdot))$，$J \in \mathbf{R}^1$. 由于 J 是一个依赖控制函数 $u(t)$ 的实数(即 J 是一个泛函)，所以有时又称 $J(u(\cdot))$ 为性能指标泛函. 通常 $J(u(\cdot))$ 可表达为

$$J(\boldsymbol{u}(\cdot)) = K(\boldsymbol{x}(t_f), t_f) + \int_{t_0}^{t_f} L(\boldsymbol{x}(t), \boldsymbol{u}((t), t)\mathrm{d}t. \tag{1-20}$$

其中 $\boldsymbol{x}(t)$ 是(1-17)式对应 $\boldsymbol{u}(t)$ 的解;$\boldsymbol{x}(t_f)$ 是 $\boldsymbol{x}(t)$ 在 t_f 时刻的值;$K(\cdots)$、$L(\cdots)$ 为其变元的标量函数.

当 $K \neq 0, L \neq 0$ 时,称 $J(\boldsymbol{u}(\cdot))$ 为混合型指标;当 $K \neq 0$,而 $L = 0$ 时,称 $J(\boldsymbol{u}(\cdot))$ 为末值型指标;当 $K = 0, L \neq 0$ 时,称 $J(\boldsymbol{u}(\cdot))$ 为积分型指标.值得注意的是:在最优控制理论中只讨论 $J(\boldsymbol{u}(\cdot)) \in \mathbf{R}^1$ 的情况.虽然可以提出多个性能指标问题,但由于多个性能指标之间无全序可比性,难于讨论.关于这方面的研究,至今,实质性的进展不大.

1.2.4 最优控制问题的数学描述及最优解

所谓**最优控制问题**系指在容许控制函数集 $\mathscr{U}$ 中找出一控制函数,使得性能指标 $J(\boldsymbol{u}(\cdot))$ 达到极小(或极大).由于使 $J(\boldsymbol{u}(\cdot))$ 达到极小的 $\boldsymbol{u}(\cdot)$ 必使 $-J(\boldsymbol{u}(\cdot))$ 达到极大,因此,从理论上讲只讨论使性能指标达到极小便足够了.使性能指标达到极小的容许控制函数称为最优控制问题(1-17)式 ~ (1-20)式的**最优控制函数**,简称**最优控制**,记为 $\boldsymbol{u}^*(t)$.(1-17)式的对应于 $\boldsymbol{u}^*(t)$ 的解 $\boldsymbol{x}^*(t)$ 称为最优控制问题(1-17)式 ~ (1-20)式的**最优轨线**.对应于 $\boldsymbol{u}^*(t)$, $\boldsymbol{x}^*(t)$ 的性能指标

$$J^* \xlongequal{\text{def}} J(\boldsymbol{u}^*) = K(\boldsymbol{x}^*(t_f^*), t_f^*) + \int_{t_0}^{t_f^*} L(\boldsymbol{x}^*(t), \boldsymbol{u}^*(t), t)\mathrm{d}t,$$

称为**最优性能指标**.如果 t_f 是不固定的,则与 $\boldsymbol{u}^*(t)$, $\boldsymbol{x}^*(t)$, J^* 相对应的 t_f^* 称为最优终端时刻;而 $t_f^* - t_0$ 称为最优过渡时间,$(\boldsymbol{x}^*(t), \boldsymbol{u}^*(t))$ 称为最优控制问题(1-17)式 ~ (1-20)式的解.

1.2.5 最优控制问题与古典变分问题的区别

众所周知,古典变分法也曾讨论过与混合型、末值型、积分型性能指标相对应的保尔茨(Paultz)、马伊尔、拉格朗日(J.L.Lagrange)变分问题.这是二者相同之处.但二者也有区别,其中最本质的区别是:在古典变分法中相当于容许控制函数的取值范围 U 是开集;而在最优控制问题中,U 可以是开集,亦可以是闭集.特别当 U 为有界闭集(它是工程上最常见的情况)时,最优控制问题的讨论就要麻烦得多.

2 极大值原理

2.1 极大值原理概述

2.1.1 庞特里亚金极大值原理

为了叙述极大值原理,对最优控制问题中所涉及的函数 $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)$, $L(\boldsymbol{x}, \boldsymbol{u},$

$t),K(x,t),g(x,t)$ 作如下假设：设 $f(x,u,t),L(x,u,t),K(x,t),g(x,t)$ 关于其变元是连续的，关于 x,t 是连续可微的，且 $f(x,u,t),\frac{\partial f}{\partial x},\frac{\partial f}{\partial t},\frac{\partial L}{\partial x},\frac{\partial L}{\partial t}$ 是有界的.

记最优控制问题的哈密顿(Hamilton)函数为

$$H(x,u,\psi,t) = -L(x,u,t) + \psi^{T}f(x,u,t). \tag{2-1}$$

定理1(庞特里亚金极大值原理) 设 $u^*(t)$ 是最优控制，$x^*(t)$ 是最优轨线，则一定存在矢值函数 $\psi(t)\in \mathbf{R}^n$ 和常矢量 $\mu\in \mathbf{R}^p$，使得在区间 $[t_0,t_f]$ 上，$u^*(t)$，$x^*(t),\psi(t),\mu$ 一起满足

$$\begin{aligned}
1^\circ\quad & \dot{x}^*(t) = \left(\frac{\partial H(x^*(t),u^*(t),\psi(t),t)}{\partial \psi}\right)^{T} = f(x^*(t),u^*(t),t),\\
& x^*(t_0) = x_0,\\
& \dot{\psi}^{T}(t) = -\frac{\partial H(x^*(t),u^*(t),\psi(t),t)}{\partial x},\\
& \dot{\psi}^{T}(t_f) = -\frac{\partial K(x^*(t_f),t_f)}{\partial x} - \mu^{T}\frac{\partial g(x^*(t_f),t_f)}{\partial x};
\end{aligned}\tag{2-2}$$

2° 对于 $u^*(t)$，在 $[t_0,t_f]$ 上的一切连续时刻 t 处，有

$$H(x^*(t),u^*(t),\psi(t),t) = \max_{u\in U}H(x^*(t),u,\psi(t),t); \tag{2-3}$$

3° 当终端时刻 t_f 不固定时，有

$$H(x^*(t),u^*(t),\psi(t),t) = H(x^*(t_f),u^*(t_f),\psi(t_f),t_f) + \int_{t_f}^{t}\frac{\partial H(x^*(t),u^*(t),\psi(t),t)}{\partial t}\mathrm{d}t, \tag{2-4}$$

$$H(x^*(t_f),u^*(t_f),\psi(t_f),t_f) = \frac{\partial K(x^*(t_f),t_f)}{\partial t_f} + \mu^{T}\frac{\partial g(x^*(t_f),t_f)}{\partial t_f}. \tag{2-5}$$

亦称最大值原理为最优控制存在的必要条件.

当终端时刻 t_f 固定时，虽然关系式(2-4)式仍成立，但此时 $H(x^*(t_f),u^*(t_f),\psi(t_f),t_f)$ 没有明确的表达式，因而提供不出更多的信息.

最优轨线 $x^*(t)$(亦称最优状态演化)和协态 $\psi(t)$ 满足的 $2n$ 阶微分方程(2-2)式，称为最优控制问题的正则方程组. 最优控制理论中，正则方程组是以 x、ψ 为未知函数并带有控制的两点边界值问题. 协态变量满足的终端点条件

$$\psi^{T}(t_f) = -\frac{\partial K(x^*(t_f),t_f)}{\partial x} - \mu^{T}\frac{\partial g(x^*(t_f),t_f)}{\partial x}$$

称为横截条件. (2-3)式是极大值原理的缘由，它的直观含义是：哈密顿函数 $H(x^*(t),u,\psi(t),t)$ 作为 u 的函数在 $u^*(t)$ 处达到极大. 由极大值原理确定出的控制称为极值控制，其相应的轨线称为极值轨线.

2.1.2 极大值原理的定解条件和最优控制综合

所谓极大值原理的定解条件系指：如果从最优控制问题的物理背景或其他途径可以知道最优控制存在且唯一，那么是否能由极大值原理确定出这个最优控制呢?答案是肯定的. 这是因为从极大值原理知，欲求最优控制 $u^*\in \mathbf{R}^m$，必须同时

求 $x^* \in \mathbf{R}^n, \psi \in \mathbf{R}^n, u^* \in \mathbf{R}^m, \mu \in \mathbf{R}^p, t_f$，共有 $2n + m + p + 1$ 个变量．而极大值原理中(2-2)式提供了 $2n$ 个条件（n 个初值条件，n 个终值条件），(2-3)式提供了 m 个条件，(2-5)式提供了一个条件，再加上 $g(x^*(t_f), t_f) = 0$ 提供的 p 个条件，共有 $2n + m + p + 1$ 个条件．它表明，在求解最优控制过程中待求的变量数目恰好等于极大值原理提供的条件数目，因而，问题是相容的．即如果已知最优控制存在且唯一，而且由极大值原理求得的解又是唯一的，则这个解就一定是最优控制问题的解；如果已知最优控制存在且唯一，但由极大值原理可求得多个解，则可通过比较其性能指标值大小的方法来获得最优控制问题的解，例如取性能指标值最小者对应的控制和相应轨线为最优的解．

如果求得的最优控制仅是时间 t 的函数，即 $u^*(t)$，则此最优控制称为程序式开环控制．如果能把最优控制解成状态 x 和时间 t 的函数，即 $u^*(t, x)$，则称它为**最优综合控制函数**．显然这种控制是状态反馈式的，因而是闭环控制．众所周知，闭环控制系统具有抗外干扰能力强和适应性强的优点．因此，在实际应用中，总是希望找到最优综合控制函数．但遗憾的是，除极少数简单的最优控制问题能求得其最优综合控制函数外，大部分最优控制问题不要说求得最优综合控制函数，甚至连开环最优控制函数的解析式都难于求得，只能通过计算机求其数字解．

2.1.3 最优控制的充分条件

由于极大值原理只是最优控制的必要条件，因此，由极大值原理解出的极值控制不一定是最优控制．所谓**最优控制充分条件**系指为使极值控制必为最优控制而应加在最优控制问题上的条件．已知最优控制的充分条件有多种形式．为了说明问题，下面给出一种．

定理 2 给定最优控制问题：

$$\dot{x} = A(t)x + f(u, t),$$
$$x(t_0) = x_0,$$
$$J(u) = C^{\mathrm{T}}x(t_f) + \int_{t_0}^{t_f}(p^{\mathrm{T}}(t)x(t) + L(u(t)))\mathrm{d}t.$$

其中 $p(t)$ 和 $A(t)$ 的元都是 t 的连续函数；$L(\cdot)$ 和 $f(\cdot)$ 都是变元的连续函数．记

$$H = -p^{\mathrm{T}}(t)x - L(u) + \psi^{\mathrm{T}}Ax + \psi^{\mathrm{T}}f(u, t),$$

设 $u^*(t) \in \mathscr{U}$（容许控制函数集），相应轨线为 $x^*(t)$．$x^*(t)$ 和 $\psi(t)$ 为方程

$$\dot{x} = A(t)x + f(u^*(t)t),$$
$$x(t_0) = x_0,$$
$$\dot{\psi}^{\mathrm{T}} = -\frac{\partial H(x, \psi, u^*(t), t)}{\partial x},$$
$$\psi^{\mathrm{T}}(t_f) = -C^{\mathrm{T}}$$

的解．如果 $u^*(t), x^*(t), \psi(t)$ 一起满足

$$H(x^*(t), u^*(t), \psi(t), t) = \max_{\hat{u} \in U} H(x^*(t), \hat{u}, \psi(t), t),$$

则 $u^*(t)$ 必为最优控制.

2.2 极大值原理与动态规划方法

2.2.1 最优性原理

1957 年贝尔曼(R.Bellman)在《动态规划》一书中提出了最优性原理.最优性原理系指:一个最优过程的任何最后一段过程都是最优的.它是一个非常一般的原理.下面以连续时间最优控制过程(最优控制问题)为例来说明这个原理.考虑如下最优控制问题.

状态方程: $\dot{x} = f(x, u), \quad x(t_0) = x_0.$ (2-6)

目标集:

控制约束: $u \in U \subset \mathbf{R}^m \quad x(t_f) = 0.$ (U 为有界闭集). (2-7)

性能指标:

$$J(u) = = \int_{t_0}^{t_f} L(x, u)\mathrm{d}t. \tag{2-8}$$

其中 $f(\cdot\cdot)$, $L(\cdot\cdot)$ 的含义和假设如 2.1.1 小节所定义.对于最优控制问题(2-6)式 ~ (2-8)式,最优性原理为:如果 $u^*(t)$ 是区间 $[t_0, t_f]$ 上的最优控制, $x^*(t)$ 是最优轨线,那么将 $u^*(t)$ 限制在 $[t_0, t_f^*]$ 内的任一子区间 $[t, t_f^*]$ 上, u^* 仍是对应于初值条件 $(t, x^*(t))$ 的最优控制.

2.2.2 动态规划的基础——贝尔曼方程

设对于每个 $x_0 \in \mathbf{R}^n$,最优控制问题(2-6)式 ~ (2-8)式的最优控制都存在,记初值条件 (t_0, x_0) 下的最优性能指标为

$$J(u^*) = J(t_0, x_0),$$

其中 $J(\cdot,\cdot)$ 是变元 t_0, x_0 的函数,且 $J(t_f, 0) = 0$.它表明对系统(2-6)式而言, t_f 时刻的状态 $x(t_f)$ 已在目标集 $J(x(t_f) = 0)$ 上,故最优性能指标为零.

记 $u^*(t)$ 为最优控制, $x^*(t)$ 为最优轨线, t_f^* 为最优终端时刻.如果作为变元 t, x 函数的最优性能指标 $J(t, x)$ 关于变元是二次连续可微的,则可得最优控制存在的另外一个必要条件,即

$$\begin{aligned}
&\min_{u \in U}\left\{\frac{\partial J(t, x^*(t))}{\partial t} + \frac{\partial J(t, x^*(t))}{\partial x} f(x^*(t), u) + L(x^*(t), u)\right\} \\
&= \frac{\partial J(t, x^*(t))}{\partial t} + \frac{\partial J(t, x^*(t))}{\partial x} f(x^*(t), u^*(t)) + \\
&L(x^*(t), u^*(t)) = 0,
\end{aligned} \tag{2-9}$$

且

$$J(t_f^*, 0) = 0. \tag{2-10}$$

关系式(2-9)式称为最优控制问题(2-6)式～(2-8)式关于性能指标 $J(\boldsymbol{u}(\cdot))$ 的**贝尔曼方程**.它是一个带边界条件(2-10)式的一阶偏微分方程.贝尔曼方程是动态规划的基础.

2.2.3 贝尔曼方程与极大值原理

通常称方程

$$\min_{\boldsymbol{u}\in U}\left\{\frac{\partial J}{\partial t}+\frac{\partial J}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u})+L(\boldsymbol{x},\boldsymbol{u})\right\}=0 \tag{2-11}$$

$$J(t_f,0)=0, \tag{2-12}$$

为连续过程的动态规划方程,即贝尔曼方程.其中 J 是 $(t,\boldsymbol{x})$ 的未知函数.如果存在一个依赖于 $\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}$ 的矢值函数 $\boldsymbol{u}=\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)$ 满足(2-11)式,即

$$\begin{aligned}&\min_{\boldsymbol{u}\in U}\left\{\frac{\partial J}{\partial t}+\frac{\partial J}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u})+L(\boldsymbol{x},\boldsymbol{u})\right\}\\&=\frac{\partial J}{\partial t}+\frac{\partial J}{\partial \boldsymbol{x}}\boldsymbol{f}\left(\boldsymbol{x},\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)\right)+L\left(\boldsymbol{x},\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)\right),\end{aligned}$$

且与 $\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)$ 相对应的偏微分方程

$$\frac{\partial J}{\partial t}+\frac{\partial J}{\partial \boldsymbol{x}}\boldsymbol{f}\left(\boldsymbol{x},\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)\right)+L\left(\boldsymbol{x},\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right)\right)=0,$$

$$J(t_f,0)=0,$$

存在关于变元 $(t,\boldsymbol{x})$ 具有二次连续可微的解 $J^*(t,\boldsymbol{x})$,则由此解 $J^*(t,\boldsymbol{x})$ 可得出最优控制问题(2-6)式～(2-8)式的相应极大值原理的全部条件.即由 $J^*(t,\boldsymbol{x})$ 可得如下函数:

$$\boldsymbol{u}(\boldsymbol{x},t)\xlongequal{\text{def}}\boldsymbol{u}\left(\boldsymbol{x},\frac{\partial J}{\partial t},\frac{\partial J}{\partial \boldsymbol{x}}\right).$$

设 $\boldsymbol{x}^*(t)$ 为如下方程的解:

$$\begin{aligned}&\dot{\boldsymbol{x}}=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u}^*(\boldsymbol{x},t)),\\&\boldsymbol{x}(t_0)=\boldsymbol{x}_0,\quad \boldsymbol{x}(t_f)=0.\end{aligned}$$

令

$$\boldsymbol{u}^*(t)\xlongequal{\text{def}}\boldsymbol{u}(\boldsymbol{x}^*(t),t),$$

取

$$\boldsymbol{\psi}^{\mathrm{T}}(t)\xlongequal{\text{def}}-\frac{\partial J^*(t,\boldsymbol{x}^*(t))}{\partial \boldsymbol{x}},$$

$$\boldsymbol{\psi}^{\mathrm{T}}(t_f)=-\frac{\partial J^*(t_f,0)}{\partial \boldsymbol{x}}=\boldsymbol{\mu}^{\mathrm{T}}.$$

其中 $\boldsymbol{\mu}$ 为待定常矢量,则 $\boldsymbol{x}^*(t),\boldsymbol{u}^*(t),\boldsymbol{\psi}(t)$ 满足相应极大值原理的全部条件.

2.2.4 最优综合控制函数存在的充分条件

给定贝尔曼方程

$$\min_{u\in U}\left\{\frac{\partial J}{\partial t}+\frac{\partial J}{\partial x}f(x,u)+L(x,u)\right\}=0, \tag{2-13}$$

$$J(t_f,x(t_f))=0. \tag{2-14}$$

其中 $L(x,u)$ 是 x,u 的正函数,即除去当 $x=0,u=0$ 时,有 $L(0,0)=0$ 外,对于其他任意 $x\in\mathbf{R}^n,u\in U$,皆有 $L(x,u)>0$.

如果存在关于变元(t,x)连续可微的矢值函数 $u(x,\frac{\partial J}{\partial t},\frac{\partial J}{\partial x})$,使得 $\frac{\partial J}{\partial t}+\frac{\partial J}{\partial x}f(x,u)+L(x,u)$ 达到极小,且偏微分方程

$$\frac{\partial J}{\partial t}+\frac{\partial J}{\partial x}f(x,u(x,\frac{\partial J}{\partial t},\frac{\partial J}{\partial x}))+L(x,u(x,\frac{\partial J}{\partial t},\frac{\partial J}{\partial x}))=0,$$

$$J(t_f,x(t_f))=0,$$

存在正定解,即存在关于 x 的正定函数 $J^*(t,x)(J^*(t,x)>0,\forall x\neq 0)$,使得 $J^*(t,x)$ 和

$$u^*(x,t)\xlongequal{\text{def}}u\left(x,\frac{\partial J^*(t,x)}{\partial t},\frac{\partial J^*(t,x)}{\partial x}\right)$$

一起满足

$$\frac{\partial J^*(t,x)}{\partial t}+\frac{\partial J^*(t,x)}{\partial x}f(x,v)+L(x,v)$$

$$\geqslant\frac{\partial J^*(t,x)}{\partial t}+\frac{\partial J^*(t,x)}{\partial x}f(x,u^*(x,t))+L(x,u^*(x,t))=0$$

$$(\forall v\in U,\forall x\in\mathbf{R}^n),$$

其中 $t_f>t_0$,则称贝尔曼方程(2-13)式、(2-14)式存在正解$(J^*(t,x),u^*(x,t))$.

定理3 如果贝尔曼方程(2-13)式、(2-14)式存在正解$(J^*(t,x),u^*(x,t))$,则 $u^*(x,t)$ 必是最优控制问题(2-6)式~(2-8)式的最优综合控制函数,而 $J^*(t,x)$ 是系统的最优性能指标.

3 时间最优控制

时间最优控制或**快速控制**通常是指以过渡时间为性能指标的最优控制.它要求把系统从一个状态(初始状态)转移到另一个状态(终端状态)的过渡时间最短.

3.1 仿射非线性系统的快速控制

仿射非线性系统的快速控制问题可描述如下.

状态方程:

$$\dot{x}=f(x,t)+B(x,t)u,\quad x(t_0)=x_0\in\mathbf{R}^n. \tag{3-1}$$

控制约束:

$$u \in U_m \xlongequal{\text{def}} \{u \mid u = (u_1, u_2, \cdots, u_m)^{\mathrm{T}}, \mid u_j \mid \leqslant 1, j = 1,2,\cdots,m\}. \tag{3-2}$$

目标集：

$$S: g(x(t_f), t_f) = 0. \tag{3-3}$$

性能指标：

$$J(u(\cdot)) = \int_{t_0}^{t_f} \mathrm{d}t = t_f - t_0. \tag{3-4}$$

其中

$$f(x,t) = (f_1(x,t), f_2(x,t), \cdots, f_n(x,t))^{\mathrm{T}},$$
$$g(x,t) = (g_1(x,t), g_2(x,t), \cdots, g_p(x,t))^{\mathrm{T}},$$
$$B(x,t) = [b_{ij}(x,t)]_{\substack{i=1,2,\cdots,n,\\ j=1,2,\cdots,m}}.$$

设 $f_i(x,t), b_{ij}(x,t), g_k(x,t), i = 1,2,\cdots,n, j = 1,2,\cdots,m, k = 1,2,\cdots,p$，都是关于其变元为连续，且关于 x 为连续可微的函数.

系统(3-1)式的快速控制问题系指求满足控制约束(3-2)式的分段连续矢值函数 $u(t)$，它把(3-1)式的初态 x_0 最快地(使指标(3-4)式达到最小)控制到目标集 S 上.

3.1.1 正则快速控制系统

如果记快速控制问题(3-1)式～(3-4)式的哈密顿函数为 $H(x,u,\psi,t)$，则有

$$H(x,u,\psi,t) = -1 + \sum_{i=1}^{n} \psi_i f_i(x,t) + \sum_{j=1}^{m} u_j \sum_{i=1}^{n} b_{ij}(x,t) \psi_i. \tag{3-5}$$

设 $u^*(t)$ 是快速控制，$x^*(t)$ 是相应的快速轨线，t_f^* 是最优终端时刻，由极大值原理可知，存在矢值函数 $\psi(t)$，使得 $x^*(t), u^*(t)$ 和 $\psi(t)$ 一起满足

$$\dot{x}^*(t) = f(x^*(t),t) + B(x^*(t),t)u^*(t),$$
$$x^*(t_0) = x_0; \tag{3-6}$$

$$\dot{\psi}_i(t) = -\sum_{l=1}^{n} \frac{\partial f_l(x^*(t),t)}{\partial x_i} \psi_l(t) - \sum_{j=1}^{m} u_j^*(t) \sum_{l=1}^{n} \frac{\partial b_{lj}(x^*(t),t)}{\partial x_i} \psi_l(t);$$
$$\psi_i(t_f^*) = -\sum_{k=1}^{p} \frac{\partial g_k(x^*(t_f^*),t_f^*)}{\partial x_i} \mu_k, \qquad i = 1,2,\cdots,n. \tag{3-7}$$

$$u_j^*(t) = \operatorname{sgn}\Big(\sum_{i=1}^{n} b_{ij}(x^*(t),t)\psi_i(t)\Big),$$
$$(\forall t \in [t_0, t_f^*]), \quad j = 1,2,\cdots,m. \tag{3-8}$$

$$-1 + \sum_{i=1}^{n} \psi_i(t_f^*) f_i(x^*(t_f^*),t_f^*) + \sum_{j=1}^{m} u_j^*(t_f^*) \sum_{l=1}^{n} b_{lj}(x^*(t_f^*),t_f^*)\psi_l(t_f^*)$$
$$= \sum_{k=1}^{p} \mu_k \frac{\partial g_k(x^*(t_f^*),t_f^*)}{\partial t_f}. \tag{3-9}$$

其中 $\mu_i, i = 1,2,\cdots,p$，是待定常量. 而

$$\operatorname{sgn}\{z\} = \begin{cases} 1 & (\text{当 } z > 0 \text{ 时}), \\ -1 & (\text{当 } z < 0 \text{ 时}), \\ ? & (\text{当 } z = 0). \end{cases}$$

显然,从(3-8)式和 sgn{·}的定义可知,为了能够通过极大值原理提供的条件确定出快速控制(如果快速控制存在),必须限制所讨论的快速控制问题为正则的.通常称它为**正则快速控制系统**.下面给出一个快速控制问题为正则的确切定义:如果对共轭方程(3-7)式的任一非零解 $\boldsymbol{\psi}(t)=(\psi_1(t),\psi_2(t),\cdots,\psi_n(t))^{\mathrm{T}}$,都不存在一个 $j(1\leqslant j\leqslant m)$ 使得函数

$$p_j(t)\xlongequal{\text{def}}\sum_{i=1}^{n}b_{ij}(\boldsymbol{x}^*(t),t)\psi_i(t)$$

在任意有限时间区间$[t_0,t_f]$上存在零聚点,则称该快速控制系统为正则的,否则称其为奇异的.值得注意的是:正则与**奇异快速控制系统**的区别仅表现在通过极大值原理提供的条件能否确定出极值控制而已,而它与快速控制问题是否存在快速控制没有什么必然的联系.也就是说,正则快速系统也可能不存在快速控制,而奇异快速系统却可能存在着快速控制,只是通过极大值原理提供的条件不能确定出它来而已.

3.1.2 正则快速控制系统的砰砰原理

正则快速控制系统表明:对于一切的 $j=1,2,\cdots,m$,$p_j(t)$ 在任何有限区间$[t_0,t_f]$上仅有有限多个零点.从(3-8)式可知,快速控制函数 $u_j^*(t),j=1,2,\cdots,m$ 都是在1和 -1 之间来回切换.切换时刻恰是 $p_j(t)$ 的零点时刻,通常称它为 $u_j^*(t),j=1,2,\cdots,m$ 的开关时刻.

定理1 (砰砰(bang-bang)原理) 设 $\boldsymbol{u}^*(t)$ 是最优控制问题(3-1)式~(3-4)式的快速控制,$\boldsymbol{x}^*(t)$ 是相应轨线,t_f^* 是相应的终端时刻,$\boldsymbol{\psi}(t)$ 是相应的共轭矢量.如果快速控制系统是正则的,则快速控制 $\boldsymbol{u}^*(t)=(u_1^*(t),u_2^*(t),\cdots,u_m^*(t))^{\mathrm{T}}$ 由(3-8)式确定.而 $\boldsymbol{x}^*(t),\boldsymbol{\psi}(t),t_f^*$ 分别满足(3-6)式、(3-7)式和(3-9)式.

3.2 线性系统的快速控制

在(3-1)式中,如果$\boldsymbol{f}(\boldsymbol{x},t)=\boldsymbol{A}(t)\boldsymbol{x}$,$\boldsymbol{B}(\boldsymbol{x},t)=\boldsymbol{B}(t)$,其中 $\boldsymbol{A}(t)$,$\boldsymbol{B}(t)$ 分别为 $n\times n$ 和 $n\times m$ 矩阵,则仿射非线性系统(3-1)式成为线性时变系统:

$$\begin{aligned}\dot{\boldsymbol{x}}&=\boldsymbol{A}(t)\boldsymbol{x}+\boldsymbol{B}(t)\boldsymbol{u},\\ \boldsymbol{x}(t_0)&=\boldsymbol{x}_0\in\mathbf{R}^n.\end{aligned}\tag{3-10}$$

设 $\boldsymbol{A}(t)$,$\boldsymbol{B}(t)$ 的元都是时间 t 的连续或分段连续函数,且记

$$\boldsymbol{B}(t)=[\boldsymbol{b}_1(t),\boldsymbol{b}_2(t),\cdots,\boldsymbol{b}_m(t)],$$

其中 $\boldsymbol{b}_j(t),j=1,2,\cdots,m$ 是$\boldsymbol{B}(t)$ 的第 j 列矢量.

3.2.1 线性时变正则快速控制系统的充分条件

线性时变快速控制问题如下.

状态方程:

$$\dot{x} = A(t)x + B(t)u,$$
$$x(t_0) = x_0.$$

控制约束:

$$|u_j| \leqslant 1, j = 1,2,\cdots,m, \quad u = (u_1, u_2, \cdots, u_m)^{\mathrm{T}}. \tag{3-11}$$

目标集:

$$x(t_f) = 0 \qquad (t_f > t_0). \tag{3-12}$$

性能指标:

$$J(u(\cdot)) = t_f - t_0. \tag{3-13}$$

所谓线性时变快速控制问题是求满足(3-11)式的分段连续控制函数 $u(t)$,它把系统(3-10)式的初态 x_0 转移到终态 $x(t_f) = 0$,且使 $t_f - t_0$ 达到极小.

易知,快速控制问题(3-10)式 ~ (3-13)式的哈密顿函数、共轭方程和横截条件分别为

$$H(x,u,\psi,t) = -1 + \psi^{\mathrm{T}}A(t)x + \psi^{\mathrm{T}}B(t)u, \tag{3-14}$$

$$\begin{aligned}\dot{\psi} &= -A^{\mathrm{T}}(t)\psi,\\ \psi(t_f) &= -\mu,\end{aligned} \tag{3-15}$$

其中 μ 为待定常矢量.

从线性微分方程的解可知,如果记 $\Phi(t,\tau)$ 为相应于矩阵 $A(t)$ 的基本解阵,则(3-15)式的共轭方程的解为

$$\psi(t) = -\Phi^{\mathrm{T}}(t_f,t)\mu, \tag{3-16}$$

从(3-14)式和(3-16)式可知,使 $H(x,u,\psi,t)$ 关于 u 取极大的 u 应为

$$u_j(t) = -\operatorname{sgn}(\mu^{\mathrm{T}}\Phi(t_f,t)b_j(t)) \quad (j = 1,2,\cdots,m). \tag{3-17}$$

显然,如果在任何有限时间区间上,函数

$$p_j(t) = \mu^{\mathrm{T}}\Phi(t_f,t)b_j(t) \quad (j = 1,2,\cdots,m)$$

仅有有限个零点,则快速控制问题(3-10)式 ~ (3-13)式是正则的.

定理 2 设 $A(t), B(t)$ 关于 t 是足够能微分的,如果对于平行坐标轴的单位矢量

$$e_j = (\underbrace{0\cdots0}_{j-1}\ 1\ \underbrace{0\cdots0}_{m-j})^{\mathrm{T}},$$

皆有

$$\operatorname{rank}[B_1(t)e_j, B_2(t)e_j, \cdots, B_n(t)e_j] = n \quad (j = 1,2,\cdots,m, \forall t \geqslant t_0),$$

则线性时变快速系统是正则的,其中

$$B_1(t) = B(t),$$

$$B_k(t) = -A(t)B_{k-1}(t) + \dot{B}_{k-1}(t) \quad (k = 2,3,\cdots,n).$$

3.2.2 线性定常正则快速控制系统的充要条件

在(3-10)式中,如果 $A(t) = A, B(t) = B, A, B$ 为常矩阵,则线性定常快速控

制问题如下.

状态方程:

$$\begin{cases}\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u};\\ \boldsymbol{x}(t_0) = \boldsymbol{x}_0.\end{cases} \tag{3-18}$$

控制约束:

$$|u_j| \leqslant 1, j = 1,2,\cdots,m, \quad \boldsymbol{u} = (u_1,u_2,\cdots,u_m)^{\mathrm{T}}. \tag{3-19}$$

目标集:

$$\boldsymbol{x}(t_f) = \boldsymbol{0}. \tag{3-20}$$

性能指标:

$$J(\boldsymbol{u}(\cdot)) = t_f - t_0. \tag{3-21}$$

所谓线性定常快速问题是求满足(3-19)式的分段连续控制函数 $\boldsymbol{u}(t)$,它把系统(3-18)式的初态 $\boldsymbol{x}_0$ 转移到终态 $\boldsymbol{x}(t_f) = 0$,且使 $t_f - t_0$ 达到极小.

对于线性定常快速控制问题,利用极大值原理可知,其快速控制函数为

$$u_j^*(t) = -\operatorname{sgn}(\boldsymbol{\mu}^{\mathrm{T}}\exp(\boldsymbol{A}^{\mathrm{T}}(t_f - t))\boldsymbol{b}_j) \quad (j = 1,2,\cdots,m), \tag{3-22}$$

其中

$$\boldsymbol{B} = [\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m].$$

定理 3　线性定常快速控制问题(3-18)式 ~ (3-21)式是正则快速控制系统的充要条件为

$$\operatorname{rank}[\boldsymbol{b}_j,\boldsymbol{A}\boldsymbol{b}_j,\cdots,\boldsymbol{A}^{n-1}\boldsymbol{b}_j] = n \quad (j = 1,2,\cdots,m),$$

称之为"**最广位置条件**".

推论 1　线性定常快速控制问题(3-18)式 ~ (3-21)式是奇异快速控制系统的充要条件为至少存在一个整数 $j_0, 1 \leqslant j_0 \leqslant m$,使得

$$\operatorname{rank}[\boldsymbol{b}_{j_0},\boldsymbol{A}\boldsymbol{b}_{j_0},\cdots,\boldsymbol{A}^{n-1}\boldsymbol{b}_{j_0}] < n.$$

定理 3 表明,线性定常快速控制问题(3-18)式 ~ (3-21)式是正则快速控制系统的充要条件,对每一个控制分量 u_j 而言,系统(3-18)式都是完全能控的.相反,(3-18)式 ~ (3-21)式是奇异快速控制系统的充要条件为至少存在一个控制分量,使得系统(3-18)式对这个控制分量而言不是完全能控的.

3.2.3　线性定常系统快速控制的唯一性

对于线性定常正则快速系统而言,由于

$$p_j(t) \xlongequal{\text{def}} \boldsymbol{\mu}^{\mathrm{T}}\exp(\boldsymbol{A}^{\mathrm{T}}(t_f - t))\boldsymbol{b}_j \qquad (j = 1,2,\cdots,m),$$

在任何有限时间区间$[t_0,t_f]$上仅有有限个零点,因此形如(3-22)式的控制函数 $\boldsymbol{u}^*(t)$ 在任何有限时间区间$[t_0,t_f]$上除去有限个时刻($p_j(t) = 0$ 的时刻)外都是完全确定的.因而,有如下定理.

定理 4　设线性定常快速控制问题(3-18)式 ~ (3-21)式是正则的.如果快速控制存在,则它必是唯一的.即是说,如果存在两个快速控制 $\boldsymbol{u}^*(t)$ 和 $\tilde{\boldsymbol{u}}^*(t)$(它们必有相同的性能指标,因此,有相同的终端时刻 t_f^*),则在$[t_0,t_f^*]$上,除有限个

时刻外，均有

$$u^*(t) = \tilde{u}^*(t).$$

对线性定常正则快速控制系统而言，如果快速控制存在，则其快速控制函数，除有限个时刻(称为开关时刻)外，都是唯一确定的.关于快速控制函数开关次数(开关时刻个数)有如下定理.

定理 5 (开关次数定理) 设线性定常快速控制问题(3-18)式～(3-21)式是正则的，且其快速控制

$$u^*(t) = (u_1^*, u_2^*, \cdots, u_m^*)^{\mathrm{T}}$$

存在，记 $u_j^*(t)$ 的开关次数为 N_j，如果矩阵 A 的特征值 $\lambda_i (i = 1,2,\cdots,n)$ 皆为实数，则有

$$N = \max_j \{N_j, j = 1,2,\cdots,m\} \leqslant n - 1,$$

其中 n 是系统(3-18)式的阶数；N 称为快速控制的开关次数.

3.2.4　线性定常系统快速控制的存在性

记线性定常快速控制问题(3-18)式～(3-21)式的容许控制函数集合为 $\mathscr{U}_m$，即

$$\mathscr{U}_m \xlongequal{\text{def}} \{u(t)\},$$

其中 $u(t)$ 为定义在有限时间区间上的取值满足约束(3-19)式的分段连续矢值函数，它把系统(3-18)式的初态 $x_0 \in \mathbf{R}^n$ 在有限时间内转移到终态 $x(t_f) = 0$.

定理 6 给定线性定常快速控制问题(3-18)式～(3-21)式，如果 $\mathscr{U}_m$ 非空，则必存在快速控制.

定理 7 设线性定常快速控制问题(3-18)式～(3-21)式是正则快速系统，如果 A 的特征值皆具有负实部，则快速控制必存在(即对于任一 $x_0 \in \mathbf{R}^n$，都存在把 x_0 引导到坐标原点的快速控制).当 $m = 1$(即系统(3-18)式为单输入)时，定理的条件可减弱为 A 的特征值皆具有非正实部.

3.2.5　双积分模型的快速控制

双积分模型是二阶线性定常系统中最简单和最常见的一个系统，其动力学方程为

$$m\ddot{y} = f(t), \tag{3-23}$$

其中 $y \in \mathbf{R}^1$，m 为常量，$f(t)$ 为驱动项.(3-23)式实际描述了一个具有惯性的质量为 m 的质点在无阻尼条件下的运动.

例如，卫星等宇航飞行器的单通道姿态控制系统常可简化成模型(3-23)式.此时 m 表示惯量矩(视为常量)，y 表示卫星绕某惯性主轴的角位移，$f(t)$ 表示控制力矩.工程上控制力矩的取值总受到限制，因此，总存在 $M > 0$ 使得 $|f(t)| \leqslant M$，M 为常数.卫星姿态控制的重要问题之一为：要求对系统施加控制力矩，使得其角位移和角速度尽快地转移到希望值上.

令 $x_1 = y, x_2 = \dot{y} = \dot{x}_1$，再通过适当选择控制变量的度量单位使其归一化，可将方程(3-23)式变换为

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = u, \\ x_1(t_0) = x_{10} = y(t_0), \quad x_2(t_0) = x_{20} = \dot{y}(t_0); \end{cases} \tag{3-24}$$

控制约束变为

$$|u| \leqslant 1; \tag{3-25}$$

终端状态为

$$x_1(t_f) = y(t_f) = 0, \quad x_2(t_f) = \dot{y}(t_f) = 0; \tag{3-26}$$

性能指标为

$$J(u(\cdot)) = t_f - t_0. \tag{3-27}$$

(3-24)式～(3-27)式组成的最优控制问题恰是线性定常快速控制问题，称为**双积分模型快速控制**.

不难验证，线性定常快速控制问题(3-24)式～(3-27)式是正则的，且其系统矩阵为

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

其中 A 有两个零特征值.

只要注意到此时 $m = 1$，从控制存在性和唯一性定理可知，把系统(3-24)式从任意初态 $(x_{10}, x_{20})^T \in \mathbf{R}^2$ 引导到坐标原点的快速控制函数存在且唯一，而且由极大值原理唯一确定，即快速控制函数为

$$u^*(t) = \operatorname{sgn}(\psi_2(t)), \tag{3-28}$$

其中共轭向量 $(\psi_1(t), \psi_2(t))^T = \boldsymbol{\psi}(t)$ 满足：

$$\begin{cases} \dot{\psi}_1 = 0, & \psi_1(t_f) = -\mu_1, \\ \dot{\psi}_2 = -\psi_1, & \psi_2(t_f) = -\mu_2. \end{cases} \tag{3-29}$$

其中 μ_1, μ_2 为待定常量.

直接解共轭方程，得

$$\psi_2(t) = -\mu_1(t_f - t) - \mu_2,$$

由上式可知，快速控制 $u^*(t)$ 至多开关一次.

注意到快速控制函数 $u^*(t)$ 的上述特点，利用时间反推法，能够得到双积分系统快速综合控制函数为

$$u^*(t, x_1, x_2) = -\operatorname{sgn}\left(x_1 + \frac{x_2}{2}|x_2|\right), \tag{3-30}$$

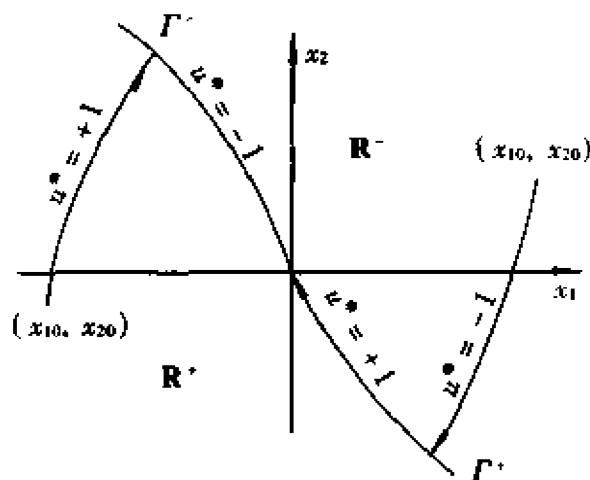
而从任一初态 $(x_{10}, x_{20})^T \in \mathbf{R}^2$ 出发，最快到达原点 $(0,0)^T$ 的过渡时间为

$$t_f^* - t_0 = \begin{cases} x_{10} + \sqrt{4x_{10} + 2x_{20}^2} & (\text{当 } x_{10} + \frac{1}{2}x_{20}|x_{20}| > 0 \text{时}); \\ -x_{10} + \sqrt{-4x_{10} + 2x_{20}^2} & (\text{当 } x_{10} + \frac{1}{2}x_{20}|x_{20}| < 0 \text{时}); \\ |x_{20}| & (\text{当 } x_{10} + \frac{1}{2}x_{20}|x_{20}| = 0 \text{时}). \end{cases} \tag{3-31}$$

将(3-30)式代入(3-24)式得双积分环节快速闭环系统

$$\begin{cases} \dot{x}_1 = x_2, & x_1(t_0) = x_{10}; \\ \dot{x}_2 = -\operatorname{sgn}(x_1 + \frac{x_2}{2}|x_2|), & x_2(t_0) = x_{20}. \end{cases} \tag{3-32}$$

(3-32)式的相平面图如图3-1所示.



$\Gamma^- = \{(x_1\ x_2) \mid x_1 = -\frac{1}{2}x_2^2, x_2 \geqslant 0\}$, $\Gamma^+ = \{(x_1\ x_2) \mid x_1 = \frac{1}{2}x_2^2, x_2 \leqslant 0\}$

图 3-1

图中 $\Gamma^- \cup \Gamma^+$ 是(3-24)式的快速闭环系统的开关曲线.而双积分环节快速闭环系统的工程实现框图如图3-2所示.

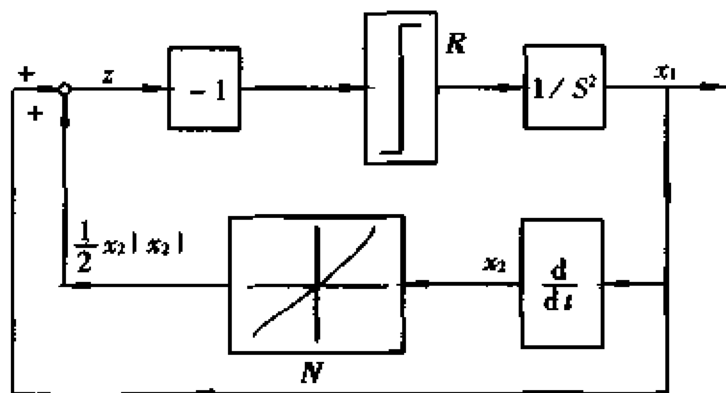


图 3-2

在图 3-2 中，R 表示继电器完成符号函数功能；N 是一非线性环节，其输入为 x_2，而输出为 $\frac{1}{2}x_2|x_2|$；$z = x_1 + \frac{1}{2}x_2|x_2|$.

4　线性二次最优控制

在实际控制工程中，存在着两类问题：① 求跟踪系统的已知"标称量"（如"标称控制函数"和"标称轨线"）或期望信号的控制设计问题. ② 求系统的误差和"控制能量"在总体平均意义下最小的控制设计问题. 这两类问题最终都归结为线性系统二次指标下的最优控制问题，简称**线性二次最优控制**问题. 从理论和应用两个方面来讲，二次最优控制问题是研究得比较完整的，其最优控制具有反馈的好形式. 特别对于线性定常系统，不但其最优控制只要通过一系列标准化过程便可求解，而且最优闭环系统也具有优良的品质，因此，深受广大工程技术人员欢迎.

4.1　线性二次最优控制问题

线性二次最优控制问题可叙述如下.

状态方程：

$$\dot{x} = A(t)x + B(t)u;\quad x(t_0) = x_0 \in \mathbf{R}^n. \tag{4-1}$$

控制约束：

$$u \in U_m, \tag{4-2}$$

其中 $U_m = \mathbf{R}^m$（或 U_m 为 $\mathbf{R}^m$ 中的开集）.

目标集：

$$S = \mathbf{R}^n \quad \text{（即终态无特别约束）}. \tag{4-3}$$

性能指标（**二次性能指标**）：

$$J(u(\cdot)) = \frac{1}{2}x^{\mathrm{T}}(t_f)Fx(t_f) + \frac{1}{2}\int_{t_0}^{t_f}[x^{\mathrm{T}}(t)Q(t)x(t) + u^{\mathrm{T}}(t)R(t)u(t)]\mathrm{d}t, \tag{4-4}$$

其中 $x \in \mathbf{R}^n, u \in \mathbf{R}^m, A(t) \in \mathbf{R}^{n\times n}, B(t) \in \mathbf{R}^{n\times m}, F \in \mathbf{R}^{n\times n}, Q(t) \in \mathbf{R}^{n\times n}, R(t) \in \mathbf{R}^{m\times m}$. 在时间区间 $[t_0, t_f]$ 上，$A(t), B(t), Q(t), R(t)$ 的元皆为 t 的连续（或分段连续）函数，且 $Q(t) \geqslant \mathbf{0}_n$（$Q(t)$ 为非负定阵），$R(t) > \mathbf{0}_m$（$R(t)$ 为正定阵），$F \geqslant \mathbf{0}_n$（F 为非负定常阵）. t_f 是事先给定的终端时刻，$\mathbf{0}_n, \mathbf{0}_m$ 分别是 $n \times n, m \times n$ 零矩阵.

线性二次最优控制问题系指，寻找定义在时间区间 $[t_0, t_f]$ 上的连续（或分段连续）控制函数 $u(t)$，它和对应的(4-1)式的解 $x(t)$ 一起使得由(4-4)式确定的 $J(u(\cdot))$ 达到极小.

由(4-1)式 ~ (4-4)式描述的线性二次最优控制问题又称线性时变系统二次最优控制问题.

4.2 线性时变系统的二次最优控制

4.2.1 有限时间最优控制的结构形式

下面讨论由(4-1)式 ~ (4-4)式描述的线性时变二次最优控制问题的解.由极大值原理可知:该系统的哈密顿函数、共轭方程和横截条件分别为

$$H(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\psi},t) = -\frac{1}{2}[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}(t)\boldsymbol{x} + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}(t)\boldsymbol{u}] + \boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{A}(t)\boldsymbol{x} + \boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{B}(t)\boldsymbol{u},$$

$$\dot{\boldsymbol{\psi}} = -\left[\frac{\partial H}{\partial \boldsymbol{x}}\right] = -\boldsymbol{A}^{\mathrm{T}}(t)\boldsymbol{\psi} + \boldsymbol{Q}(t)\boldsymbol{x},$$

$$\boldsymbol{\psi}(t_f) = -\boldsymbol{F}\boldsymbol{x}(t_f). \tag{4-5}$$

使 $H(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\psi},t)$ 取得极大的 $\boldsymbol{u}$ 应有如下形式:

$$\boldsymbol{u}(t,\boldsymbol{\psi}) = \boldsymbol{R}^{-1}(t)\boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\psi}. \tag{4-6}$$

将(4-6)式代入(4-1)式联合(4-5)式解得相应的两点边界值问题:

$$\begin{cases} \begin{bmatrix} \dot{\boldsymbol{x}} \\ \dot{\boldsymbol{\psi}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(t) & \boldsymbol{B}(t)\boldsymbol{R}^{-1}(t)\boldsymbol{B}^{\mathrm{T}}(t) \\ \boldsymbol{Q}(t) & -\boldsymbol{A}^{\mathrm{T}}(t) \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\psi} \end{bmatrix}; \\ \boldsymbol{x}(t_0) = \boldsymbol{x}_0, \\ \boldsymbol{\psi}(t_f) = -\boldsymbol{F}\boldsymbol{x}(t_f). \end{cases} \tag{4-7}$$

(4-7)式的解 $(\boldsymbol{x}^{\mathrm{T}}(t),\boldsymbol{\psi}^{\mathrm{T}}(t))^{\mathrm{T}}$ 可表达为

$$\boldsymbol{\psi}(t) = -\boldsymbol{P}(t)\boldsymbol{x}(t),$$

其中 $\boldsymbol{P}(t) \in \mathbf{R}^{n\times n}$,且满足带终端条件的里卡蒂(J.F.Riccati)**矩阵微分方程**:

$$\dot{\boldsymbol{P}} + \boldsymbol{P}\boldsymbol{A}(t) + \boldsymbol{A}^{\mathrm{T}}(t)\boldsymbol{P} + \boldsymbol{Q}(t) - \boldsymbol{P}\boldsymbol{B}(t)\boldsymbol{R}^{-1}(t)\boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{P} = \boldsymbol{0}_n,$$

$$\boldsymbol{P}(t,t_f) = \boldsymbol{F} \qquad (t \in [t_0,t_f]). \tag{4-8}$$

由关于 $\boldsymbol{A}(t),\boldsymbol{B}(t),\boldsymbol{Q}(t),\boldsymbol{R}(t)$ 的假定可知,里卡蒂矩阵微分方程(4-8)式的解 $\boldsymbol{P}(t)$ 在时间区间 $[t_0,t_f]$ 上存在且唯一,而且具有性质 $\boldsymbol{P}^{\mathrm{T}}(t) = \boldsymbol{P}(t)$.因此,线性时变二次最优控制问题的解 $(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t))$ 应有如下形式:

$$\boldsymbol{u}^*(t) = -\boldsymbol{R}^{-1}(t)\boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{P}(t)\boldsymbol{x}^*(t), \tag{4-9}$$

即如果线性时变二次最优控制问题的最优控制 $\boldsymbol{u}^*(t)$ 存在,则它和最优轨线 $\boldsymbol{x}^*(t)$ 必须满足(4-9)式,它是有限时间最优控制的结构形式.

4.2.2 有限时间最优控制的存在性

为了指出有限时间线性时变二次最优控制问题(4-1)式 ~ (4-4)式的最优控制存在,利用里卡蒂矩阵微分方程(4-8)式的解 $\boldsymbol{P}(t)$,构造控制函数:

$$\boldsymbol{u}(t,\boldsymbol{x}) = -\boldsymbol{R}^{-1}(t)\boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{P}(t)\boldsymbol{x}. \tag{4-10}$$

将(4-10)式代人(4-1)式,得

$$\begin{aligned}\dot{x} &= (A(t) - B(t)R^{-1}(t)B^{\mathrm{T}}(t)P(t))x;\\ x(t_0) &= x_0.\end{aligned} \tag{4-11}$$

记(4-11)式的解为 $x^*(t)$.通过考查二次型函数 $x^{*\mathrm{T}}(t)P(t)x^*(t)$,利用配方方法可证明:

$$u^*(t) = -R^{-1}(t)B^{\mathrm{T}}(t)P(t)x^*(t),$$

这恰是有限时间线性时变二次最优控制问题(4-1)式 ~ (4-4)式的最优控制.由此,结合4.2.1小节的内容可得到如下定理和推论.

定理1 对于给定的有限时间线性时变二次最优控制问题(4-1)式 ~ (4-4)式,其最优控制存在且唯一.其表达式为

$$u^* = -R^{-1}(t)B^{\mathrm{T}}(t)P(t)x,$$

其中 $P(t)$ 是里卡蒂矩阵微分方程(4-8)式的唯一对称解,而最优性能指标为

$$J^* = J(u^*(t)) = \frac{1}{2}x_0^{\mathrm{T}}P(t_0)x_0 \geqslant 0. \tag{4-12}$$

推论1 由(4-12)式且注意到 t_0 的任意性可知,里卡蒂矩阵微分方程(4-8)式的解 $P(t)$,在时间区间 $[t_0, t_f]$ 上是非负对称阵.即

$$P(t) \geqslant 0 \quad (\forall t \in [t_0, t_f]).$$

由定理1可知,状态反馈形式的控制函数(4-10)式恰是有限时间线性时变二次最优控制问题(4-1)式 ~ (4-4)式的最优综合控制函数.工程上通常称它为最优控制器.

4.2.3 无穷时间最优控制的结构形式

如果(4-1)式 ~ (4-4)式中的 $A(t)$,$B(t)$,$Q(t)$,$R(t)$ 的元在无穷区间 $[t_0, +\infty]$ 上连续,且 $Q(t) \geqslant 0_n$,$R(t) \geqslant R_0$,$R_0 > 0_m$,即 R_0 是一个 $m \times m$ 正定常矩阵,在(4-1)式 ~ (4-4)式中取 $t_f \to \infty$,且 $F = 0$,则可得到线性时变无穷区间 $[t_0, +\infty)$ 上的二次最优控制问题.除性能指标变为

$$J(u(\cdot)) = \frac{1}{2}\int_{t_0}^{+\infty}[x^{\mathrm{T}}(t)Q(t)x(t) + u^{\mathrm{T}}(t)R(t)u(t)]\mathrm{d}t \tag{4-13}$$

外,状态方程、控制约束、目标集均与(4-1)式 ~ (4-3)式的相同.由于性能指标(4-13)式中的积分区间为 $[t_0, +\infty)$,因此,线性时变无穷区间二次最优控制问题的容许控制函数集合应为

$$\mathscr{U}[t_0, +\infty) \xlongequal{\text{def}} \{u(t)\},$$

其中 $u(t)$ 为取值满足(4-2)式的在 $[t_0, +\infty)$ 上为分段连续的矢值函数,且 $u(t)$ 和其对应的(4-1)式的初值问题的解 $x(t)$,使得(4-13)式的 $J(u(\cdot))$ 为有限.其最优控制问题则是在 $\mathscr{U}[t_0, +\infty)$ 中选一控制函数使 $J(u(\cdot))$ 达到最小.

有两种方式可对已知线性时变无穷区间二次最优控制问题进行讨论.① 直接讨论;② 将有限区间线性时变二次最优控制问题的解通过 $t_f \to +\infty$ 的极限来获得.下面采取后一种方式进行讨论.

从有限时间线性时变二次最优控制问题(4-1)式～(4-4)式的最优控制的解析形式可知,里卡蒂矩阵微分方程的解 $P(t,t_f)$ 依赖于终端时刻 t_f,因此,必须讨论 $\lim\limits_{t_f\to+\infty} P(t,t_f)$ 的情况.关于 $\lim\limits_{t_f\to+\infty} P(t,t_f)$ 有如下引理.

引理1　对于任给定的 t_f,设 $P(t,t_f)$ 是里卡蒂矩阵微分方程(4-8)式关于终端条件 $P(t_f,t_f)=\mathbf{0}_n$ 的唯一非负定解.如果对于每一个 $t\in[t_0,+\infty)$,系统(4-1)式都是完全能控的,则有

$$\lim_{t_f\to+\infty} P(t,t_f)=\bar{P}(t)\quad(\forall t\geqslant t_0).\tag{4-14}$$

极限矩阵 $\bar{P}(t)$ 为非负定阵,且满足

$$\begin{aligned}\dot{\bar{P}}(t)+\bar{P}(t)A(t)+A^{\mathrm{T}}(t)\bar{P}(t)+Q(t)-\\ \bar{P}(t)B(t)R^{-1}(t)B^{\mathrm{T}}(t)\bar{P}(t)=\mathbf{0}_n\quad(\forall t\geqslant t_0).\end{aligned}\tag{4-15}$$

定理2　对于给定的线性时变无穷区间二次最优控制问题(4-1)式～(4-3)式及(4-13)式,如果对于每一个 $t\in[t_0,+\infty)$,系统(4-1)式都是完全能控的,则线性时变无穷区间二次最优控制问题(4-1)式～(4-3)式及(4-13)式的解存在且唯一,且最优综合控制函数为

$$u^*(t,x)=-R^{-1}(t)B^{\mathrm{T}}(t)\bar{P}(t)x,\tag{4-16}$$

其中 $\bar{P}(t)$ 是由(4-14)式确定的极限矩阵.

工程上,通常称(4-16)式中的 $u^*(t,x)$ 为线性时变系统的二次最优调节器,而

$$G(t)\xlongequal{\text{def}}R^{-1}(t)B^{\mathrm{T}}(t)\bar{P}(t)$$

称为**二次最优反馈增益矩阵**.显然二次最优闭环系统为

$$\begin{aligned}\dot{x}&=(A(t)-B(t)R^{-1}(t)B^{\mathrm{T}}(t)\bar{P}(t))x,\\ x(t_0)&=x_0.\end{aligned}$$

4.3　线性定常系统的二次最优控制

4.3.1　有限时间的最优综合控制函数

如果在(4-1)式～(4-4)式中,$A(t)=A$,$B(t)=B$,$Q(t)=Q\geqslant\mathbf{0}_n$,$R(t)=R>\mathbf{0}_m$,且 A,B,Q,R 为相应维数的常阵,则此时与(4-1)式～(4-4)式相应的控制问题称为有限时间线性定常二次最优控制问题.这个问题的结论,只有定理1一个定理表述它,即其最优综合控制函数存在且唯一,其表达式为

$$u^*(t,x)=-R^{-1}B^{\mathrm{T}}P(t,t_f)x,$$

其中 $P(t,t_f)$ 是里卡蒂矩阵微分方程

$$\begin{aligned}\dot{P}+PA+A^{\mathrm{T}}P+Q-PBR^{-1}B^{\mathrm{T}}P&=\mathbf{0}_n,\\ P(t,t_f)&=F\end{aligned}\tag{4-17}$$

的唯一非负定解,其最优性能指标为

$$J^* = \frac{1}{2}x_0^{\mathrm{T}}P(t,t_f)x_0. \tag{4-18}$$

值得注意的是，虽然线性定常系统比线性时变系统简单，但是有限时间线性定常二次最优控制问题的最优综合函数对应的反馈阵仍是时变的. 只不过其里卡蒂矩阵微分方程的求解可能会简单一些.

例 1　给定有限时间线性定常二次最优控制问题

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$

$$x(0) = \begin{bmatrix} 3 \\ 2 \end{bmatrix},$$

$$J(u(\cdot)) = \frac{1}{2}x^{\mathrm{T}}(3)\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}x(3) + \frac{1}{2}\int_0^3\left(x^{\mathrm{T}}(t)\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}x(t) + \frac{1}{2}u^2(t)\right)\mathrm{d}t.$$

求其最优综合控制函数.

解　据定理 1 可知，若记 $P(t,3) = \begin{bmatrix} p_{11}(t,3) & p_{12}(t,3) \\ p_{12}(t,3) & p_{22}(t,3) \end{bmatrix}$，$x = (x_1, x_2)^{\mathrm{T}}$，则

$$\begin{aligned} u^*(t,x) &= -R^{-1}B^{\mathrm{T}}P(t,3)x = -2[0\ 1]P(t,3)x \\ &= -2(p_{12}(t,3)x_1 + p_{22}(t,3)x_2), \end{aligned}$$

而 $P(t,3)$ 满足：

$$\dot{P}(t,3) + P(t,3)\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}P(t,3) + \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} -$$

$$P(t,3)\begin{bmatrix} 0 \\ 1 \end{bmatrix} 2[0\ 1]P(t,3) = 0_n,$$

$$P(3,3) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

展开上式得

$$\dot{p}_{11}(t,3) = 2p_{12}^2(t,13) - 2, \qquad p_{11}(3,3) = 1,$$

$$\dot{p}_{12}(t,3) = -p_{11}(t,3) + 2p_{12}(t,3)p_{22}(t,3) - 1, \quad p_{12}(3,3) = 0,$$

$$\dot{p}_{22}(t,3) = -2p_{12}(t,3) + 2p_{22}^2(t,3) - 4, \quad p_{22}(3,3) = 2.$$

从上述方程组中解出 $p_{11}(t,3)$，$p_{12}(t,3)$，$p_{22}(t,3)$，再代入到 $u^*(t,x)$ 的表达式中，便可得其最优综合控制函数. 由于 $p_{11}(t,3)$，$p_{12}(t,3)$，$p_{22}(t,3)$ 都是 t 的函数，从而知其反馈增益阵是时变的.

4.3.2　无穷时间的最优调节

在(4-1)式 ~ (4-4)式中，假定

$A(t) = A$，$B(t) = B$，$Q(t) = Q \geqslant 0_n$，$R(t) = R > 0_m$.

其中 A，B，Q，R 皆为常阵.

若在(4-1)式 ~ (4-4)式中取 $t_f \to +\infty$，且 $F = 0$，则可得无穷时间线性定常二次最优控制问题：

$$\dot{x} = Ax + Bu, \tag{4-19}$$
$$x(t_0) = x_0,$$
$$u \in U_m = \mathbf{R}^m, \tag{4-20}$$
$$J(u(\cdot)) = \frac{1}{2}\int_0^{+\infty}(x^{\mathrm{T}}(t)Qx(t) + u^{\mathrm{T}}(t)Ru(t))\mathrm{d}t, \tag{4-21}$$

工程上通常称它为线性二次最优调节问题.

引理 2　对于任意给定的 t_f,设 $P(t,t_f)$ 是终端条件为 $P(t_f,t_f) = 0_n$ 的里卡蒂矩阵微分方程(4-17)式的唯一非负定解.如果系统(4-19)式是完全能控的,即

$$\mathrm{rank}[B, AB, \cdots, A^{n-1}B] = n,$$

则有

$$\lim_{t_f \to +\infty} P(t,t_f) = P^* \quad (\text{常阵}). \tag{4-22}$$

其中极限矩阵 P^* 为**里卡蒂矩阵代数方程**.

$$PA + A^{\mathrm{T}}P + Q - PBR^{-1}B^{\mathrm{T}}P = 0 \tag{4-23}$$

的非负定对称解.

定理 3　对于给定的线性二次最优调节问题(4-19)式 ~ (4-21)式,如果系统(4-19)式是完全能控的,则线性二次最优调节问题(4-19)式 ~ (4-21)式的解存在且唯一.其最优综合控制函数为

$$u^*(x) = -R^{-1}B^{\mathrm{T}}P^*x, \tag{4-24}$$

其中 P^* 为**里卡蒂矩阵代数方程**(4-23)式的非负定对称解.

工程上称形如(4-24)式的最优综合控制函数为**最优调节器**.此时,最优反馈增益阵

$$G^* = R^{-1}B^{\mathrm{T}}P^*$$

是一个常阵.而最优闭环系统为

$$\dot{x} = (A - BR^{-1}B^{\mathrm{T}}P^*)x, \tag{4-25}$$
$$x(t_0) = x_0.$$

例 2　假定有一线性二次最优调节问题:

$$\dot{x} = x + u,$$
$$x(t_0) = x_0,$$
$$J(u(\cdot)) = \frac{1}{2}\int_{t_0}^{\infty}(x^2 + u^2)\mathrm{d}t.$$

求最优综合控制函数,并讨论最优闭环系统的稳定性.

解　在该系统中,$A = 1, B = 1, Q = 1, R = 1$,则里卡蒂代数方程为

$$p^* + p^* + 1 - p^{*2} = 0,$$

上述二次方程的解为

$$p^* = 1 \pm \sqrt{2}.$$

由于 $p^* \geqslant 0$,所以　$p^* = 1 + \sqrt{2}$.该系统的最优综合控制函数为

$$u^*(x) = -R^{-1}B^{\mathrm{T}}p^*x = -(1+\sqrt{2})x.$$

将其代入原系统，得最优闭环系统为

$$\dot{x} = x + u^*(x) = (1-1-\sqrt{2})x = -\sqrt{2}x,$$
$$x(t_0) = x_0.$$

其解为

$$x^*(t) = x_0\exp(-\sqrt{2}(t-t_0)).$$

显然，当 $t\to\infty$ 时，有 $\lim\limits_{t\to\infty}x^*(t) = 0$，即最优闭环系统是渐近稳定的.

例3 假定有线性最优调节问题：

$$\dot{x} = x + u,$$
$$x(t_0) = x_0,$$
$$J(u(\cdot)) = \frac{1}{2}\int_{t_0}^{\infty} u^2\mathrm{d}t.$$

求最优综合控制函数，并讨论其最优闭环系统的稳定性.

解 在该系统中，$A = 1, B = 1, Q = 0, R = 1$，则里卡蒂代数方程为

$$p^* + p^* - p^{*2} = 0,$$

即

$$p^*(2-p^*) = 0,$$

所以

$$p_1^* = 0,\quad p_2^* = 2.$$

p_1^* 对应的最优综合控制函数为

$$u_1^*(x) = 0.$$

p_2^* 对应的最优综合控制函数为

$$u_2^*(x) = -2x.$$

显然

$$0 = J(u_1^*) < J(u_2^*),$$

因此，其最优综合控制函数为

$$u^*(x) = 0,$$

而最优闭环系统为

$$\dot{x}^* = x^*,$$
$$x^*(t_0) = x_0.$$

其解为 $x^*(t) = x_0\exp(t-t_0)$.

由上述的解可知，其最优闭环系统是不稳的.

两个例子的最优闭环解的稳定性性质相反，两个例子的不同之处在于性能指标中的加权阵 Q 不同.因此，为了使最优闭环系统是渐近稳定的，必须对性能指标中的矩阵 Q 做某些假定.

4.3.3 线性定常二次最优闭环系统的稳定性

设$[A,C^{\mathrm{T}}]$是完全能观的，即

$$\operatorname{rank}\begin{bmatrix} C^{T} \\ C^{T}A^{T} \\ \vdots \\ C^{T}(A^{T})^{n-1} \end{bmatrix} = n,$$

其中 C^{T} 是使

$$Q = C\,C^{T}$$

成立的任一矩阵.为了说明上述假设有意义,必须证明上述假设成立且与 Q 的分解无关.

命题 1　设 Q 有两组不同分解

$$Q = C\,C^{T} = C_1\,C_1^{T},$$

则$[A, C^{T}]$完全能观的充要条件是$[A, C_1^{T}]$完全能观.

命题 2　对于给定的线性二次最优调节问题(4-19)式 ~ (4-21)式,设系统(4-19)式完全能控.P^{*} 是里卡蒂矩阵代数方程

$$PA + A^{T}P + Q - PBR^{-1}B^{T}P = 0$$

的解.则 $P^{*} > 0_n$ 的充要条件为,对于 Q 的任一分解 $Q = C\,C^{T}$,$[A, C^{T}]$是完全能观的.

定理 4　对于给定的线性定常二次最优调节问题(4-19)式 ~ (4-21)式,设$[A, B]$完全能控,且对于 Q 的任一分解 $Q = C\,C^{T}$,$[A, C^{T}]$完全能观,则最优闭环系统

$$\dot{x} = (A - BR^{-1}B^{T}P^{*})x,$$

$$x(t_0) = x_0,$$

必是渐近稳定的.其中 P^{*} 是里卡蒂矩阵代数方程(4-23)式的唯一正定对称解矩阵.

例 4　对于给定的线性定常二次最优调节问题:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$

$$x(t_0) = x_0,$$

$$J(u(\cdot)) = \frac{1}{2}\int_0^{\infty}\left(x^{T}(t)\begin{bmatrix} 1 & b \\ b & a \end{bmatrix} x(t) + u^2(t)\right)\mathrm{d}t,$$

其中 a, b 为满足 $a > 0, a - b^2 > 0$ 的常数,试求最优综合控制函数,并讨论最优闭环系统的稳定性.

解　因 $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$,则 $\operatorname{rank}[B\ A\,B] = \operatorname{rank}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = 2$,再由 $Q = \begin{bmatrix} 1 & b \\ b & a \end{bmatrix}$ 和 $a - b^2 > 0$ 可知,$Q > 0$,

因此,对于 Q 的任一分解 $Q = C\,C^{T}$,必有$[A, C^{T}]$完全能观,因此该系统的最优控制存在且唯一.如果记

$$P^{*} = \begin{bmatrix} p_{11}^{*} & p_{12}^{*} \\ p_{12}^{*} & p_{22}^{*} \end{bmatrix},$$

则最优综合控制函数为

$$u^*(x) = -R^{-1}B^{\mathrm{T}}P^*x = -p_{12}^*x_1 - p_{22}^*x_2,$$

其中 P^* 是相应的里卡蒂矩阵代数方程的唯一正定解. 由

$$p_{12}^{*2} = 1,$$

$$-p_{11}^* + p_{12}^*p_{22}^* - b = 0,$$

$$-2p_{12}^* + (p_{22}^*)^2 - a = 0$$

得,

$$p_{12}^* = \pm 1,$$

$$p_{11}^* = p_{12}^*p_{22}^* - b,$$

$$p_{22}^* = \pm\sqrt{a + 2p_{12}^*}.$$

由于 p^* 是 2×2 正定对称矩阵, 即

$$p_{11}^* > 0,\quad p_{11}^*p_{22}^* - p_{12}^{*2} > 0,$$

从而得 $p_{22}^* > 0$, 因此, 有

$$p_{22}^* = \sqrt{a + 2p_{12}^*}.$$

下面讨论 p_{12}^* 的符号. 如果 $p_{12}^* = -1$, 则

$$p_{22}^* = \sqrt{a-2}.$$

为了保证 p_{22}^* 为正实数, 必有 $a > 2$. 将 $p_{12}^* = -1$ 和 $p_{22}^* = \sqrt{a-2}$ 代入到 p_{11}^* 的表达式中, 得

$$p_{11}^* = p_{12}^*p_{22}^* - b = -\sqrt{a-2} - b.$$

由于 $p_{11}^* > 0$, 从上式得

$$b < -\sqrt{a-2}.$$

将 $p_{11}^* = p_{12}^*p_{22}^* - b$ 代入 $p_{11}^*p_{22}^* - p_{12}^{*2} > 0$ 中, 得

$$p_{12}^*p_{22}^{*2} - bp_{22}^* > p_{12}^{*2}.$$

将 $p_{12}^* = -1, p_{22}^* = \sqrt{a-2}$, 代入上式, 得

$$-(a-2) - b\sqrt{a-2} > 1,$$

即

$$-b > \frac{a-1}{\sqrt{a-2}}.$$

只要注意到 $a > 2$, 则可从上式直接得

$$b^2 > \frac{(a-1)^2}{a-2} = \frac{(a-2+1)^2}{a-2} = a + \frac{1}{a-2} > a.$$

它与 $b^2 < a$ 矛盾. 于是得

$$p_{12}^* = 1,$$

$$p_{22}^* = \sqrt{a+2},$$

$$p_{11}^* = \sqrt{a+2} - b.$$

将其代入最优综合控制函数的表达式中, 有

$$u^*(x) = -x_1 - \sqrt{a+2}x_2.$$

而最优闭环系统为

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -\sqrt{a+2} \end{bmatrix} x.$$

由于$\begin{bmatrix} 0 & 1 \\ -1 & -\sqrt{a+2} \end{bmatrix}$的特征值为 $-\frac{\sqrt{a+2}}{2} \pm \frac{\sqrt{2-a}}{2}\mathrm{j}$,其中 $\mathrm{j} = \sqrt{-1}$,从而可知,最优闭环系统是渐近稳定的.

4.4 里卡蒂矩阵代数方程的求解

求解里卡蒂矩阵代数方程(4-23)式的方法很多.这里仅介绍一种利用李雅普诺夫(Lyapunov)矩阵代数方程的解序列,通过求极限得到(4-23)式的正定对称解的方法.

4.4.1 李雅普诺夫矩阵代数方程及其解

给定 $n \times n$ 矩阵 $\boldsymbol{A}, \boldsymbol{Q}$,其中 $\boldsymbol{Q}$ 为对称阵,以矩阵 $\boldsymbol{P}$ 为未知量的方程

$$\boldsymbol{PA} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{P} = -\boldsymbol{Q} \tag{4-26}$$

称为李雅普诺夫矩阵代数方程.

命题 3 对于任给的 $\boldsymbol{Q} \geqslant \boldsymbol{0}_n$ 及其分解 $\boldsymbol{Q} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$.若$[\boldsymbol{A}, \boldsymbol{C}^{\mathrm{T}}]$完全能观,则(4-26)式有解 $\boldsymbol{P} > \boldsymbol{0}_n$ 的充要条件是 $\boldsymbol{A}$ 为稳定阵(即 $\boldsymbol{A}$ 的特征值皆具有负实部),且

$$\boldsymbol{P} = \int_0^{+\infty} \exp(\boldsymbol{A}^{\mathrm{T}}t)\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}\exp(\boldsymbol{A}t)\mathrm{d}t. \tag{4-27}$$

推论 2 对于任给的 $\boldsymbol{Q} > 0$,(4-26)式有解 $\boldsymbol{P} > 0$ 的充要条件是 $\boldsymbol{A}$ 为稳定阵.

4.4.2 里卡蒂矩阵代数方程的逼近解

将里卡蒂矩阵代数方程(4-23)式改写为

$$\boldsymbol{P}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}) + (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P})^{\mathrm{T}}\boldsymbol{P} = -(\boldsymbol{Q} + \boldsymbol{P}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}),$$

则有如下定理.

定理 5 对于给定的里卡蒂矩阵代数方程(4-23)式,设$[\boldsymbol{A}, \boldsymbol{B}]$完全能控,且对于 $\boldsymbol{Q}$ 的任一分解 $\boldsymbol{Q} = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$,$[\boldsymbol{A}, \boldsymbol{C}^{\mathrm{T}}]$完全能观,则按迭代矩阵方程组

$$\begin{aligned}
&\boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{P}_k + \boldsymbol{P}_k\boldsymbol{A}_k = -\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k, \\
&\boldsymbol{A}_k = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_k, \\
&\boldsymbol{K}_k = \boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_{k-1}, \\
&\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k = \boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{K}_k^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_k \quad (k = 0,1,2,\cdots),
\end{aligned}$$

能够求得方程(4-23)式的唯一正定对称解阵 $\boldsymbol{P}^*$,即

$$\lim_{k\to+\infty} \boldsymbol{P}_k = \boldsymbol{P}^*,$$

且 $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}^*$ 为稳定阵.其中 $\boldsymbol{K}_0$ 为使 $\boldsymbol{A}_0 \xlongequal{\mathrm{def}} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_0$ 稳定的任一 $m \times n$ 阵;

P_0 是李雅普诺夫矩阵代数方程　$PA_0 + A_0^{\mathrm{T}}P = -D_0^{\mathrm{T}}D_0$ 的唯一正定对称解.

注 1　矩阵序列 $\{P_k, k = 1,2,\cdots\}$ 称为里卡蒂矩阵代数方程(4-23)式的逼近解序列, P_k 称为(4-23)式的第 k 次逼近解.

注 2　(4-23)式的逼近解序列 $\{P_k\}$ 具有如下性质:

1° $P_k > 0_n$　　　$(\forall k = 0,1,2,\cdots)$,

2° $P_k \leqslant P_{k-1}$　　　$(\forall k = 1,2,\cdots)$,

3° $P^* \leqslant P_k$　　　$(\forall k = 0,1,2,\cdots)$,

其中 P^* 是(4-23)式的唯一正定解阵.

4.5　具有指定衰减度的二次最优调节

4.5.1　线性定常齐次方程解的衰减度

给定线性定常齐次矢量常微分方程:

$$\begin{aligned}\dot{x} &= Fx;\\ x(t_0) &= x_0,\end{aligned} \tag{4-28}$$

其解记为 $x(t, t_0, x_0)$. 如果 F 为稳定阵,则一定存在一个正数 $\alpha > 0$,使得

$$\lim_{t\to\infty} x(t, t_0, x_0)\exp(\alpha t) = 0,$$

则称 α 为(4-28)式的衰减度. 对于自由系统(4-28)式,其衰减度不唯一,而且也不能事先指定;对于受控系统(4-19)式,在一定条件下,选择适当的线性状态反馈形式的控制,可使相应闭环系统具有事先指定衰减度.

4.5.2　具有指定衰减度的线性二次最优调节问题

对于任给定的正数 β,考查如下的线性二次最优调节问题:

$$\dot{x} = Ax + Bu, \tag{4-29}$$

$$\begin{aligned}x(0) &= x_0;\\ u \in U_m &= \mathbf{R}^m,\end{aligned} \tag{4-30}$$

$$J_\beta[u(\cdot)] = \frac{1}{2}\int_0^{+\infty}(x^{\mathrm{T}}(t)Qx(t) + u^{\mathrm{T}}(t)Ru(t))\exp(2\beta t)\mathrm{d}t, \tag{4-31}$$

其中 $x \in \mathbf{R}^n, u \in \mathbf{R}^m; A, B, Q, R$ 为适当维数的常阵,且 $Q \geqslant 0_n, R > 0_m$.

设 $[A, B]$ 完全能控,对于 Q 的任一分解 $Q = CC^{\mathrm{T}}$, $[A, C^{\mathrm{T}}]$ 完全能观.

取变量变换

$$\begin{aligned}\xi(t) &= x(t)\exp(\beta t),\\ v(t) &= u(t)\exp(\beta t),\end{aligned}$$

则(4-29)式、(4-30)式和(4-31)式分别变为

$$\dot{\xi} = [\beta I_n + A]\xi + Bv, \tag{4-32}$$

$$\xi(0) = x(0) = x_0,$$

$$J_\beta(\nu(\cdot)) = \frac{1}{2}\int_0^\infty (\xi^T(t)Q\xi(t) + \nu^T(t)R\nu(t))dt, \tag{4-33}$$

$$\nu \in U_m = \mathbf{R}^m. \tag{4-34}$$

注意到$[A + \beta I_n, B]$完全能控的充要条件为$[A, B]$完全能控,$[A + \beta I_n, C^T]$完全能观的充要条件为$[A, C^T]$完全能观,则对于线性二次最优调节问题(4-32)式~(4-34)式,应用定理3可知,其最优调节器为

$$\nu^*(t) = -R^{-1}B^TP_\beta^*\xi(t),$$

而 P_β^* 是里卡蒂矩阵代数方程

$$P(A + \beta I_n) + (A + \beta I_n)^TP + Q - PBR^{-1}B^TP = 0 \tag{4-35}$$

的唯一正定对称解阵,且最优闭环系统

$$\dot{\xi} = (A + \beta I_n - BR^{-1}B^TP_\beta^*)\xi$$

是渐近稳定的,从而有

$$\lim_{t\to\infty}\xi(t) = \lim_{t\to\infty}x(t)\exp(\beta t) = 0_n.$$

由此可知,$x(t)$具有事先指定的衰减度β.

由 $\nu(t) = -R^{-1}B^TP_\beta^*\xi(t)$ 可直接得

$$u^*(t) = -R^{-1}B^TP_\beta^*x(t).$$

定理6 给定线性二次最优调节问题(4-29)式~(4-31)式.若$[A, B]$完全能控,且对于Q的任一分解$Q = CC^T$,$[A, C^T]$完全能观,则具有事先指定的衰减度β的最优调节器为

$$u^* = -R^{-1}B^TP_\beta^*x.$$

P_β^* 是里卡蒂矩阵代数方程(4-35)式的唯一正定对称解阵,而最优闭环系统的解具有事先指定的衰减度β.

4.6 线性定常系统二次最优调节的逆问题

4.6.1 二次最优调节逆问题

给定一个线性定常系统

$$\dot{x} = Ax + Bu, \tag{4-36}$$

$$x(0) = x_0.$$

所谓系统(4-36)式的**二次最优调节逆问题**系指:对于给定系统(4-36)式的一个线性状态反馈控制律

$$u = -Kx,$$

其中$K \in \mathbf{R}^{m\times n}$,是否存在或何时存在$Q_1 \in \mathbf{R}^{n\times n}$,$Q_1 \geq 0_n$和$R_1 \in \mathbf{R}^{m\times m}$,$R_1 > 0_m$,使得 $u = -Kx$ 恰是使由 Q_1 和 R_1 构成的二次性能指标

$$J_1(u(\cdot)) = \frac{1}{2}\int_0^{+\infty} (x^T(t)Q_1x(t) + u(t)R_1u(t))dt$$

达到最小的最优调节器?

4.6.2　线性二次最优调节问题的频域条件

从 4.3 节知道,求最优调节器的关键是,解里卡蒂矩阵代数方程

$$PA + A^{\mathrm{T}}P + Q - PBR^{-1}B^{\mathrm{T}}P = 0_n. \tag{4-37}$$

(4-37) 式正定解 P^* 的存在性,以及它在线性二次最优调节问题讨论中的作用,前面已给出. 下面进一步给出由(4-37) 式导出的结果.

设 P^* 是(4-37) 式的唯一正定解,令 s 表示复变量,将 P^* 代入(4-37) 式,整理后可得

$$P^*(sI_n - A) + (-sI_n - A^{\mathrm{T}})P^* + P^*BR^{-1}B^{\mathrm{T}}P^* = Q.$$

已知最优反馈增益阵 $K^* = R^{-1}B^{\mathrm{T}}P^*$,即 $K^{*\mathrm{T}}R^{\frac{1}{2}} = P^*BR^{-\frac{1}{2}}$. 若令 $s = \mathrm{j}\omega$, $\mathrm{j} = \sqrt{-1}$,由上式直接得

$$[I_m + R^{\frac{1}{2}}K^*(-\mathrm{j}\omega I_n - A)^{-1}BR^{-\frac{1}{2}}]^{\mathrm{T}}[I_m + R^{\frac{1}{2}}K^*(\mathrm{j}\omega I_n - A)^{-1}BR^{-\frac{1}{2}}]$$

$$= I_m + R^{-\frac{1}{2}}B^{\mathrm{T}}(-\mathrm{j}\omega I_n - A^{\mathrm{T}})^{-1}Q(\mathrm{j}\omega I_n - A)^{-1}BR^{-\frac{1}{2}}.$$

注意到 $Q \geqslant 0_n$,则由上式直接可得

$$[I_m + R^{\frac{1}{2}}K^{*\mathrm{T}}(-\mathrm{j}\omega I_n - A^{\mathrm{T}})^{-1}BR^{-\frac{1}{2}}]^{\mathrm{T}}[I_m + R^{\frac{1}{2}}K^{*\mathrm{T}}(\mathrm{j}\omega I_n - A)^{-1}BR^{-\frac{1}{2}}] \geqslant I_m. \tag{4-38}$$

不等式(4-38) 式称为线性二次**最优调节器频域条件**. 显然,对于任意给定的使得 $(A - BK)$ 成为稳定阵的 $m \times n$ 阶矩阵 K,(4-38) 式不一定成立.

当 $m = 1$ 时,$B = b$,不失一般性取 $R = 1$,则(4-38) 式变为

$$|1 + k^{\mathrm{T}}(\mathrm{j}\omega - A)^{-1}b| \geqslant 1, \tag{4-39}$$

上式等号仅在有限个 ω 上成立.

4.6.3　二次最优调节逆问题的解

在二次最优调节问题中,由于 $R > 0_m$,故不失一般性可取 $R = I_m$.

定理 7　对于给定的线性定常系统(4-36)式,设$[A, B]$完全能控. 对于任给的 $K_1 \in \mathbf{R}^{m\times n}$,如果 K_1 满足:

(1) $A - BK_1$ 为稳定阵,

(2) $[A, K_1]$ 完全能观测,

(3) $[I_m + K_1(-\mathrm{j}\omega I_n - A)^{-1}B]^{\mathrm{T}}[I_m + K_1(\mathrm{j}\omega I_n - A)^{-1}B] \geqslant I_m \quad (\forall\, \omega)$,

且等号仅在有限个 ω 上成立,则一定存在非负定阵 Q_1,使得 $u = -K_1x$ 恰是使二次性能指标

$$J(u(\cdot)) = \frac{1}{2}\int_0^{\infty}(x^{\mathrm{T}}(t)Q_1x(t) + u^{\mathrm{T}}(t)u(t))\mathrm{d}t$$

达到最小的最优调节器.

当 $m = 1$ 时,$B = b$,Q_1 可由如下步骤求得:

(1) 通过线性变换将(4-36) 式化成能控标准形,即

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

(2) 取常矢量 $k^T = [k_1, k_2, \cdots, k_n]$ 使 $[A - bk^T]$ 为稳定阵，且 $[A, k]$ 完全能观.

(3) 计算 A 和 $A - bk^T$ 的特征多项式：

$$\psi(s) = \det[sI_n - A] = s^n + a_n s^{n-1} + \cdots + a_2 s + a_1,$$

$$p(s) = \det[sI_n - A + bk^T] = s^n + (a_n - k_n)s^{n-1} + \cdots + (a_2 - k_2)s + (a_1 - k_1).$$

(4) 对 $p(-j\omega)p(j\omega) - \psi(-j\omega)\psi(j\omega)$ 作因式分解，得

$$p(-j\omega)p(j\omega) - \psi(-j\omega)\psi(j\omega) = l(-j\omega)l(j\omega),$$

其中

$$l(s) = d_n s^{n-1} + d_{n-1}s^{n-2} + \cdots + d_2 s + d_1,$$

则 $d^T = (d_1, d_2, \cdots, d_n) \neq 0$，且 $[A, d]$ 完全能观.

(5) $Q_1 \xlongequal{\text{def}} dd^T \geqslant 0_n$ 为所要求的非负定阵.

5 微分对策——双方极值控制

5.1 微分对策问题

虽然微分对策问题最初来源于军事上的需求，但慢慢地发现将工程和经济领域中某些与控制有关的问题视为微分对策问题更为恰当，因此，微分对策方法的应用范围不断扩大，其理论也不断发展和完善.

5.1.1 微分对策问题实例

设用全向推力火箭 D 追击目标 M，追击在平面上进行，并引用如下符号：

x_1, x_2——火箭 D 的位置坐标；

x_3, x_4——火箭 D 的速度分量；

x_5, x_6——目标 M 的位置坐标；

F——火箭推力大小，设为常数；

u——火箭推力方向与垂线的夹角；

W——目标 M 的速度，设为常数；

v——M 的速度方向与垂线的夹角；

K——空气阻力系数，设为常数.

D 和 M 的运动方程分别为

$$
\begin{cases}
\dot{x}_1 = x_3, \\
\dot{x}_2 = x_4, \\
\dot{x}_3 = F\sin u - Kx_3, \\
\dot{x}_4 = F\cos u - Kx_4, \\
\dot{x}_5 = W\sin v, \\
\dot{x}_6 = W\cos v,
\end{cases}
\tag{5-1}
$$

其中 u,v 分别是 D,M 的"控制".由于火箭的推力方向和目标的速度方向可任意,故 u,v 的取值不受限制.

当D和M之间的距离满足$(x_1 - x_5)^2 + (x_2 - x_6)^2 \leqslant l^2$时,就认为实现了"捕获".其中 l 是一个事先给定的正数.微分对策问题是:对于 D,选择 u,使实现捕获的时间 t_f 最小;而对于 M,选择 v,使实现捕获的时间 t_f 最大.

5.1.2　微分对策问题

在前面介绍的最优控制问题中,都只存在一方控制,要求控制使系统从初始状态转移到另一个所要求的终端状态,并使性能指标达到最小.然而在实际工程中,特别是在作战中,常常会出现非常复杂的情况.例如在空战中,不仅甲方飞机要施加控制,力求用火力击中乙方飞机,而乙方飞机也会施加控制,以躲避免遭击中,若时机合适,还会还击甲方飞机.这样就形成了一类双方控制问题.在这类双方控制问题中,双方施加控制的目标是相互对立的.例如当甲方飞机处在攻击态势时,甲方飞机就要施加控制,力求使双方相对距离缩小,以便更准确地击中目标;而处在躲避态势的乙方飞机,则会施加控制,力求使双方的相对距离扩大,便于跑掉.总之,在这类双方控制问题中,双方施加控制的目的是对立的,都是要使自己处于有利的地位.另外,由于在实际工程系统中总存在各种各样的不确定性,例如,系统参数的不确定性,未建模动态乃至外干扰的不确定性等,一般来说,系统的不确定性是不能确切知道的,但通过物理试验或理论分析能知道不确定性取值的一个大致范围,因此,有时可以把系统中的不确定性看成是系统中的"附加控制".它的存在阻碍系统原来目标的实现.这样,一个带有不确定性的控制系统中,就会存在控制目的相互对立的两种控制:一种是原系统中真正的控制,一种是把系统的不确定性看成另一种附加控制的控制.此时就可以把带有不确定性的控制系统考虑成为一个**微分对策问题**.

5.1.3　微分对策研究中的两类基本问题

微分对策研究中有两类基本问题:一类是二人零和性质的定量微分对策问题.其特征是,对策过程中的参与双方(亦称局中人)可通过选择各自控制,在满足共同的终端状态条件下,使某个性能指标具有对立的极值性质.定量微分对策问题和前面讨论的最优控制问题在形式上有很多相似之处,在一定意义上来说,定量微分

对策问题是最优控制问题的直接推广.另一类是定性微分对策问题.这类对策不是研究局中人选择各自的控制,使某一性能指标具有对立的极值性质,而是研究具有对抗性过程的某种结局是否能实现.如雷达监视对策:P,E两方,各具有一定的运动性能;P上装有雷达,有一定的搜索半径,P欲追踪监视E,E欲摆脱监视.今规定,若P,E间的距离大于或等于一个指定数 d,即 $|PE| \geqslant d$ 等,则E方摆脱监视,此时对策结束.E方力争摆脱,P方则相反.问在什么条件下对策结束?

5.2 一类定量微分对策

5.2.1 定量微分对策问题的数学描述

设P,E为对策双方,其状态方程为

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t); \\ \boldsymbol{x}(t_0) &= \boldsymbol{x}_0. \end{aligned} \tag{5-2}$$

其中 $\boldsymbol{x} \in \mathbf{R}^n$ 是由P,E双方形成的状态矢量(例如可取P,E双方的相对位置矢量、相对速度矢量);$\boldsymbol{u} \in \mathbf{R}^r$ 是P方的控制;$\boldsymbol{v} \in \mathbf{R}^p$ 是E方控制.通常 $r,p \leqslant n$,且

$$\boldsymbol{u} \in U, \quad \boldsymbol{v} \in V. \tag{5-3}$$

这里 $U \subset \mathbf{R}^r$,$V \subset \mathbf{R}^p$. U,V 可以是,而且常常是有界闭集.$(\boldsymbol{u},\boldsymbol{v})$ 称为策略,而 $U \times V$ 称为策略集.记为 $(\boldsymbol{u},\boldsymbol{v}) \in U \times V$. $\boldsymbol{f}(\cdot\cdot\cdot\cdot)$ 是 $n+r+p+1$ 个变元的矢值函数.通常要求它对于任一容许策略 $(\boldsymbol{u}(t),\boldsymbol{v}(t))$,都保证初值问题(5-2)式的解存在且唯一.

性能指标通常取为

$$J(\boldsymbol{u}(\cdot),\boldsymbol{v}(\cdot)) = F(\boldsymbol{x}(t_f),t_f) + \int_{t_0}^{t_f} L(\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{v}(t),t)\mathrm{d}t, \tag{5-4}$$

其中 $F(\cdot\cdot)$ 为 $n+1$ 个变元的二次可微标量函数;$L(\cdot\cdot\cdot\cdot)$ 是 $n+p+r+1$ 个变元的标量函数;t_f 是双方的策略结束的时刻,可以是事先给定的,亦可以是自由的(待定的).此外,在 $t=t_f$ 时通常要求有终端状态约束:

$$\boldsymbol{g}(\boldsymbol{x}(t_f),t_f) = \boldsymbol{0}_q, \tag{5-5}$$

其中 $\boldsymbol{g}(\cdot\cdot)$ 是 $n+1$ 个变元的二次可微 q 维矢值函数.

记容许策略集为 $\mathscr{U} \times \mathscr{V} = \{(\boldsymbol{u},\boldsymbol{v})\}$,其中 $(\boldsymbol{u}(t),\boldsymbol{v}(t))$ 为在 $[t_0,t_f]$ 上有定义,且取值于 $U \times V$ 上的分段连续 r,p 维矢值函数,它使得(5-2)式的初值问题的唯一解 $\boldsymbol{x}(t)$ 在 $t=t_f$ 时满足(5-5)式.

定量微分对策(5-2)式~(5-5)式的最优策略问题是在容许策略集 $\mathscr{U} \times \mathscr{V}$ 中选择 $(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$,对于P方,它使 $J(\boldsymbol{u}(\cdot),\boldsymbol{v}(\cdot))$ 达最小,而对于E方,它使 $J(\boldsymbol{u}(\cdot),\boldsymbol{v}(\cdot))$ 达极大,即要求 $(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$ 满足:

$$\begin{gathered} J(\boldsymbol{u}^*(\cdot),\boldsymbol{v}(\cdot)) \leqslant J(\boldsymbol{u}^*(\cdot),\boldsymbol{v}^*(\cdot)) \leqslant J(\boldsymbol{u}(\cdot),\boldsymbol{v}^*(\cdot)) \\ (\forall(\boldsymbol{u}(t),\boldsymbol{v}(t)) \in \mathscr{U} \times \mathscr{V}). \end{gathered} \tag{5-6}$$

如果满足(5-6)式的 $(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$ 存在,则称它为最优策略,又称它为鞍

点.对应于$(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$的(5-2)式的解称为最优轨线,简记为$\boldsymbol{x}^*(t)$. $J^* \stackrel{\text{def}}{=} J(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$称为最优策略指标值.与最优策略$(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$对应的终端时刻$t_f^*$,称为最优终端时刻. $(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$简称为**定量微分对策问题**(5-2)式 ~ (5-5)式的最优解.

5.2.2　定量双方极值原理

在叙述定量微分对策问题(5-2)式 ~ (5-5)式的定量双方极值原理之前,先对(5-2)式 ~ (5-5)式中涉及到的函数$\boldsymbol{f}(\cdot\cdot\cdot\cdot)$, $F(\cdot\cdot)$, $L(\cdot\cdot\cdot\cdot)$, $\boldsymbol{g}(\cdot\cdot)$作如下假定:

(1) 关于$\boldsymbol{f} = [f_1,f_2,\cdots,f_n]^{\mathrm{T}}$和$L(\cdot\cdot\cdot\cdot)$,

① $f_i(\boldsymbol{x},\boldsymbol{u},\boldsymbol{u},t)$, $i = 1,2,\cdots,n$,和$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t)$关于变元是连续的.

② $f_i(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t)$和$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t)$关于$x_j$和$t$有直到二阶的连续偏导数,即$\dfrac{\partial f_i}{\partial x_j}$, $\dfrac{\partial f_i}{\partial t}$, $\dfrac{\partial L}{\partial x_j}$, $\dfrac{\partial L}{\partial t}$, $\dfrac{\partial^2 f_i}{\partial x_k \partial x_j}$, $\dfrac{\partial^2 f_i}{\partial t \partial x_j}$, $\dfrac{\partial^2 L}{\partial x_j \partial x_k}$, $\dfrac{\partial^2 L}{\partial t \partial x_j}$都是变元的连续函数.

③ $f_i(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t)$和$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t)$关于$\boldsymbol{u},\boldsymbol{v}$满足利普希茨(R. Lipschitz)条件,即

$$
\begin{aligned}
&| f_i(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t) - f_i(\boldsymbol{x}',\boldsymbol{u}',\boldsymbol{v}',t) | \leqslant M \| \boldsymbol{u} - \dot{\boldsymbol{u}} \| + \\
&\qquad N \| \boldsymbol{v} - \boldsymbol{v}' \| \qquad\qquad (i = 1,2,\cdots,n), \\
&| L(\boldsymbol{x},\dot{\boldsymbol{u}},\dot{\boldsymbol{v}},t) - L(\boldsymbol{x}',\boldsymbol{u}',\boldsymbol{v}',t) | \leqslant M \| \boldsymbol{u} - \boldsymbol{u} \| + \\
&N \| \boldsymbol{v} - \boldsymbol{v}' \|.
\end{aligned}
$$

(2) $F(\cdot\cdot)$和$\boldsymbol{g}(\cdot\cdot)$都是关于变元有直到二阶连续偏导数.记

$$
H(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},\boldsymbol{\psi},t) = -L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t) + \boldsymbol{\psi}^{\mathrm{T}}(t)\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},t) \tag{5-7}
$$

$H(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},\boldsymbol{\psi},t)$称为定量微分对策问题(5-2)式 ~ (5-5)式的哈密顿函数.

定理1(定量双方极值原理)　设$(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t)) \in \mathscr{U} \times \mathscr{V}$, $\boldsymbol{x}^*(t)$是(5-2)式相应于$(\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$的轨线.为了使$(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t),\boldsymbol{v}^*(t))$是定量微分对策问题(5-2)式 ~ (5-5)式的最优解,必存在矢值函数$\boldsymbol{\psi}(t) \in \mathbf{R}^n$,使得$\boldsymbol{x}^*(t)$, $\boldsymbol{u}^*(t)$, $\boldsymbol{v}^*(t)$、$\boldsymbol{\psi}(t)$一起满足

$$
\begin{aligned}
1^\circ\ & \dot{\boldsymbol{x}}^*(t) = \left(\frac{\partial H}{\partial \boldsymbol{\psi}}\right)_*^{\mathrm{T}} = \boldsymbol{f}(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t),\boldsymbol{v}^*(t),t), \\
& \boldsymbol{x}^*(t_0) = \boldsymbol{x}_0, \qquad t \in [t_0, t_f^*], \\
& \boldsymbol{\psi}^{\mathrm{T}}(t) = -\left(\frac{\partial H}{\partial \boldsymbol{x}}\right)_*, \\
& \boldsymbol{\psi}^{\mathrm{T}}(t_f^*) = -\left(\frac{\partial F}{\partial \boldsymbol{x}}\right)_* - \boldsymbol{\mu}^{\mathrm{T}}\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\right)_*,
\end{aligned}
$$

其中$(\quad)_*$表示用带上标 * 号的量代入后的矢量或矩阵.

2° 在$\boldsymbol{u}^*(t)$, $\boldsymbol{v}(t)$的一切连续时刻处,有

$$
H(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t),\boldsymbol{v}^*(t),\boldsymbol{\psi}(t),t) = \max_{\boldsymbol{u}\in U}\min_{\boldsymbol{v}\in V} H(\boldsymbol{x}^*(t),\boldsymbol{u},\boldsymbol{v},\boldsymbol{\psi}(t),t)
$$

$$= \min_{v \in V} \max_{u \in U} H(x^*(t), u, v, \psi(t), t).$$

3° 哈密顿函数沿着$(\psi(t), x^*(t), u^*(t), v^*(t))$，有

$$H(x^*(t), u^*(t), v^*(t), \psi(t), t)$$
$$= H(x^*(t_f^*), u^*(t_f^*), v^*(t_f^*), \psi(t_f^*), t_f^*) + \int_{t_f}^{t} \frac{\partial H(x^*(t), u^*(t), \psi(t), t)}{\partial t} dt,$$

当 t_f 自由时，有

$$H(x^*(t_f^*), u^*(t_f^*), v^*(t_f^*), \psi(t_f^*, t_f^*) = \left(\frac{\partial F}{\partial t_f}\right)_* + \mu^{\mathrm{T}}\left(\frac{\partial g}{\partial t_f}\right)_*,$$

显然，当系统为定常系统，且 t_f 自由时，有

$$H(x^*(t), u^*(t), v^*(t), \quad \psi(t)) = 0;$$

当系统为定常系统，且 t_f 为事先给定时，则有

$$H(x^*(t), u^*(t), v^*(t), \psi(t)) = \text{常量}.$$

$\psi(t)$ 称为共轭(伴随)向量，它所满足的方程称为共轭方程，它满足的终端条件称为横截条件.

通常称

$$\begin{cases} \dot{x} = \left(\dfrac{\partial H}{\partial \psi}\right)^{\mathrm{T}}, \\ \dot{\psi} = -\left(\dfrac{\partial H}{\partial x}\right)^{\mathrm{T}} \end{cases}$$

为定量微分对策问题(5-2)式～(5-5)式的哈密顿正则方程.

例 制定导弹拦截目标的最优策略.

解 假设将导弹(下标为 D)和目标(下标为 M)都视为质点. 在某个惯性坐标系中，导弹和目标的位置矢量分别为 x_D 和 x_M，其速度矢量分别为 $v_D = \dot{x}_D, v_M = \dot{x}_M$，导弹拦截目标的示意图如图 5-1 所示.

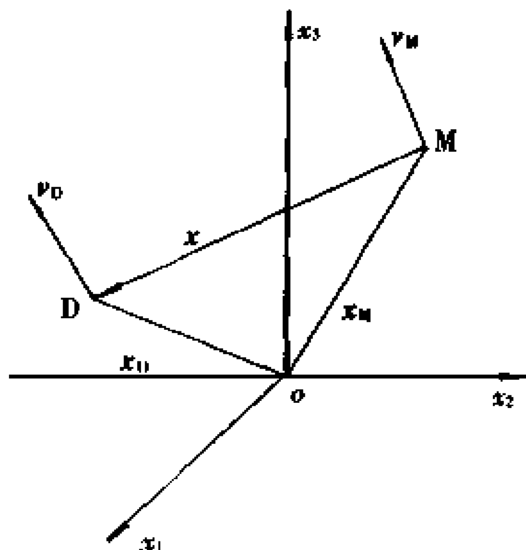


图 5-1

记 $x = x_D - x_M, v = v_D - v_M$.

在略去作用到导弹和目标上的重力差和气动力差的条件下，导弹和目标的相对运动方程为

$$\begin{cases} \dot{x} = v, \\ \dot{v} = u_D - u_M, \\ x(t_0) = x_0, \\ v(t_0) = v_0. \end{cases} \tag{5-8}$$

其中 u_D 为导弹的控制加速度，u_M 为目标的控制加速度.

若取性能指标为

$$J(u_D(\cdot),u_M(\cdot))=\frac{k}{2}x^T(t_f)x(t_f)+\frac{1}{2}\int_{t_0}^{t_f}(C_D u_D^T u_D-C_M u_M^T u_M)dt,\quad(5\text{-}9)$$

其中 k, C_D 和 C_M 都是加权常数；t_f 是事先给定的.

假设

$$u_D\in\mathbf{R}^3,\quad u_M\in\mathbf{R}^3,\quad(5\text{-}10)$$

即 u_D 和 u_M 取值大小都不受约束，则(5-8)式～(5-10)式便组成了一个定量微分对策问题.由定量双方极值原理可知，该系统的哈密顿函数为

$$H=-\frac{1}{2}C_D u_D^T u_D+\frac{C_M}{2}u_M^T u_M+\psi_1^T v+\psi_2^T(u_D-u_M).$$

其共轭方程和横截条件为

$$\dot{\psi}_1=-\left(\frac{\partial H}{\partial x}\right)^T=0,$$

$$\dot{\psi}_2=-\left(\frac{\partial H}{\partial v}\right)^T=-\psi_1,$$

$$\psi_1(t_f)=-kx(t_f),$$

$$\psi_2(t_f)=0.$$

对以上各式积分，得

$$\begin{cases}\psi_1(t)=\psi_1(t_f)=-kx(t_f),\\ \psi_2(t)=-kx(t_f)(t_f-t).\end{cases}\quad(5\text{-}11)$$

显然，当 $u_D^*(t)=\frac{1}{C_D}\psi_2(t)$ 时，H 达到极大，而当 $u_M^*(t)=\frac{1}{C_M}\psi_2(t)$ 时，H 达到极小.因此，由(5-11)式可知，其最优策略为

$$u_D^*(t)=\frac{1}{C_D}\psi_2(t)=-\frac{k}{C_D}x(t_f)(t_f-t),$$

$$u_M^*(t)=\frac{1}{C_M}\psi_2(t)=-\frac{k}{C_M}x(t_f)(t_f-t).$$

将上式代入(5-8)式，积分后，令 $t=t_f$，得

$$x(t_f)=\frac{x_0+v_0(t_f-t_0)}{1+\frac{k(C_M-C_D)}{3C_M C_D}(t_f-t_0)^3},$$

因此，有

$$\begin{cases}u_D^*(t_0)=-\dfrac{(x_0+v_0(t_f-t_0))(t_f-t_0)}{C_D\left(\dfrac{1}{k}+\dfrac{(C_M-C_D)}{3C_M C_D}(t_f-t_0)^3\right)},\\ u_M^*(t_0)=-\dfrac{(x_0+v_0(t_f-t_0))(t_f-t_0)}{C_M\left(\dfrac{1}{k}+\dfrac{(C_M-C_D)}{3C_M C_D}(t_f-t_0)^3\right)}.\end{cases}\quad(5\text{-}12)$$

由于 t_0 是任意的，所以只要 $t_0<t_f$，由(5-12)式给出的 $u_D^*(t_0)$，$u_M^*(t_0)$ 就是 t_0 时

刻的最优策略,即鞍点.

下面主要讨论 $u_D^*(t_0)$ 的具体形式. 当 $k \to +\infty$(相当于 $x(t_f)=0$) 时, 由(5-12) 式得

$$u_D^*(t_0) = -3\frac{x_0 + v_0(t_f - t_0)}{\left(1 - \dfrac{C_D}{C_M}\right)(t_f - t_0)^2}. \tag{5-13}$$

如果选择 $t_f - t_0 = -\dfrac{\langle x_0, v_0\rangle}{\langle x_0, x_0\rangle}$,则(5-13) 式变为

$$u_D^*(t_0) = \frac{3}{1 - \dfrac{C_D}{C_M}}\frac{\langle x_0, v_0\rangle}{\langle x_0, x_0\rangle}\omega_0 \times x_0, \tag{5-14}$$

其中 $\omega_0 = \dfrac{x_0 \times v_0}{\langle x_0, x_0\rangle}$ 称为视线角速度.

如果选择 $t_f - t_0 = -\dfrac{\langle x_0, x_0\rangle}{\langle x_0, v_0\rangle}$,则(5-13) 式变为

$$u_D^*(t_0) = \frac{3}{1 - \dfrac{C_D}{C_M}}\frac{|x_0|^2 |v_0|^2}{\langle x_0, v_0\rangle^2}\omega_0 \times v_0. \tag{5-15}$$

当 $C_M \to +\infty$,即在性能指标中不考虑 $u_M(t)$ 的影响,相当于目标无机动时,(5-14) 式、(5-15) 式分别变为

$$u_D^*(t_0) = 3\frac{\langle x_0, v_0\rangle}{\langle x_0, x_0\rangle}\omega_0 \times x_0,$$

$$u_D^*(t_0) = 3\frac{|x_0|^2 |v_0|^2}{\langle x_0, v_0\rangle^2}\omega_0 \times v_0.$$

前者称为变系数真比例导航规律,后者称为变系数纯比例导航规律.

5.3 一类定性微分对策

5.3.1 定性微分对策问题

设对策 P,E 双方的相对运动方程为

$$\begin{aligned}\dot{x} &= f(x, u, v);\\ x(t_0) &= x_0.\end{aligned} \tag{5-16}$$

关于方程和 u, v 的性质如 5.1.1 小节所述. 只是(5-16) 式为定常系统.

给定目标集 $$S: \psi(x) \leqslant 0_l, \tag{5-17}$$

其中 $\psi \in \mathbf{R}^l$,即 $\psi(\cdot)$ 是变元的 l 维矢值函数,且 $l \leqslant n$. P方选择 $u \in U$,力求实现(5-17) 式;E 方选择 $v \in V$,阻碍(5-17) 式的实现.

记容许策略集为

$$\mathscr{U}_P \times \mathscr{V}_E = \{(u(t), v(t)\},$$

其中 $u(t)$ 和 $v(t)$ 都是定义在有限区间上的分别取值于 U 和 V 的分段连续 r 和 p 维矢值函数.

如果存在 $(u(t),v(t))\in\mathscr{U}_P\times\mathscr{V}_E$，使得(5-16)式的解 $x(t)$ 在某时刻 $t_1\geqslant t_0$ 之后，永远满足

$$\psi(x(t))\leqslant 0\qquad(t\geqslant t_1),$$

则称双方的策略过程结束(或追赶成功，或躲避成功).

定性微分对策问题系指：在什么条件下对策结束?即在容许对策集合 $\mathscr{V}_P\times\mathscr{V}_E$ 中寻找使(5-16)式的解满足不等式(5-17)式的策略 $(u(t),v(t))$.

笼统地提问上述定性微分对策问题(5-16)式、(5-17)式在什么条件下双方的策略结束是没有意义的.这是因为若对策双方的"能力"相差悬殊，则或者对策必定能结束(不管 E 方取任何的 $v(t)$)，或者对策必定不能结束(不管 P 方取任何的 $u(t)$).此时，若是前者，那么使对策结束的策略 (u,v) 就会有无穷多个，因此，答案是不能确定的.显然，为了使答案是明确的，只能考虑双方"能力"具有一定均势的定性微分对策问题.在这种情况下，在状态空间可能存在两个区域，暂称为"捕捉区" D_P 和"躲避区" D_E.当(5-16)式的状态 $x\in D_P$ 时，无论 E 方取何种策略 $v(t)\in\mathscr{V}$，P 方总可选适当策略 $u\in\mathscr{U}$，使对策结束；当 $x\in D_E$ 时，只要 E 方取适当策略 $v\in\mathscr{V}$，则对策总不能结束.因此，具有实际意义的问题是寻找区域 D_P 和 D_E 的分界面，称为"界栅"(barier).正是在"界栅"上，双方进行最激烈的对抗，均施展其最优策略，不容半点疏忽.若双方势均力敌，则对抗就总是在"界栅"上进行，即(5-16)式的状态总在界栅上演化，于是"界栅"将由(5-16)式的轨线组成.

定性微分对策(5-16)式、(5-17)式的最优策略问题是：在 $\mathscr{U}_P\times\mathscr{V}_E$ 中寻找 $(u(t),v(t))$，它使(5-16)式的轨线 $x^*(t)$ 在某时刻 $t_1\geqslant t_0$ 后，保持在"界栅"上运动.使(5-16)式的轨线保持在"界栅"上运动的策略称为定性微分对策问题(5-16)式、(5-17)式的最优策略，并记为 $(u^*(t),v(t))$，相应的轨线记为 $x^*(t)$.

5.3.2　定性双方极值原理

由于具有双方控制特点的定性微分对策问题是以将初态 x_0 引导到一个给定的目标集 S 为对策结束的条件，所以可将它与控制理论中的能控性问题相对应.由此，可得出关于与界栅对应的最优策略所应满足的定性双方极值原理.

记 $f=(f_1,f_2,\cdots,f_n)^T$，$\psi=(\psi_1,\psi_2,\cdots,\psi_l)^T$.设

1° $f_i(x,u,v)$，$i=1,2,\cdots,n$，关于变元是连续的，关于 x_j 有直到二阶连续的偏导数，即 $\dfrac{\partial f_i}{\partial x_j}$，$\dfrac{\partial^2 f_i}{\partial x_k\partial x_j}$ 都是变元的连续函数.

2° $f_i(x,u,v)$ 关于变元 u,v 满足利普希茨条件，即

$$|f_i(x,u',v')-f_i(x,u,v)|\leqslant M\|u-u'\|+N\|v-v'\|.$$

3° $\psi_j(x)$，$j=1,2,\cdots,l$，关于变元 x_k，$k=1,2,\cdots,n$，有直到二阶的连续导数.

令 $H_1(x,u,v,\psi)=\psi^T f(x,u,v)$.

定理 2（定性双方极值原理）　设$(u^*(t), v^*(t)) \in \mathscr{U}_P \times \mathscr{V}_E$，相应(5-16)式的轨线为$x^*(t)$．为了使$(u^*(t), v^*(t))$是定性微分对策问题(5-16)式、(5-17)式的最优策略，必存在一矢值函数$\psi(t) \in \mathbf{R}^n$，它和$x^*(t)$、$u^*(t)$，$v^*(t)$一起满足

1°
$$\dot{x}^*(t) = \left(\frac{\partial H_1}{\partial \psi}\right)^{\mathrm{T}} = f(x^*(t), u^*(t), v^*(t)),$$
$$x^*(t_0) = x_0,$$
$$\dot{\psi}^{\mathrm{T}}(t) = -\left(\frac{\partial H_1}{\partial x}\right)_* \quad (t \in [t_0, t_1]),$$
$$\psi^{\mathrm{T}}(t_1) = -\mu^{\mathrm{T}}\left(\frac{\partial \psi}{\partial x}\right)_*.$$

2° 对于在$[t_0, t_1]$上$(u^*(t), v^*(t))$的一切连续的时刻，皆有
$$H_1(x^*(t), u^*(t), v^*(t), \psi(t)) = \max_{u \in U} \min_{v \in V} H_1(x^*(t), u, v, \psi(t)) = \min_{v \in V} \max_{u \in U} H_1(x^*(t), u, v, \psi(t)).$$

3° $H_1(x^*(t), u^*(t), v^*(t), \psi(t)) = 0 \quad (\forall t \in [t_0, t_1])$.

其中t为$(u^*(t), v^*(t))$的连续的时刻；t_1为以x_0为初态沿着“界栅”运动的最优轨线到达目标集S的边界的时刻，它是需要确定的变量．

5.4　斯蒂克贝格策略

斯蒂克贝格(L. Stickelberger)策略是一种主、从微分对策．对策局中人是主、从两方．主方(leader)宣布一种策略，从方(follower)响应并使自己的性能指标达到最小．在从方响应的情况下，主方再选择使自己的性能指标达到最小的策略．

5.4.1　斯蒂克贝格策略的数学描述

设主、从双方的状态方程为
$$\begin{aligned} \dot{x} &= f(x, u, v, t); \\ x(t_0) &= x_0 \in \mathbf{R}^n. \end{aligned} \tag{5-18}$$
其中$x \in \mathbf{R}^n$是状态；$u \in U$，是主方策略；$v \in V$是从方策略；U，V分别是主、从两方的策略取值集，它们通常为$\mathbf{R}^r$，$\mathbf{R}^p$中的有界闭集；$f(\cdot\cdot\cdot\cdot)$是变元的n维矢值函数，它要保证方程(5-18)式有唯一解．

在对策结束时刻t_f时，状态满足的终端条件，即目标集为
$$S: \psi(x(t_f), t_f) = 0, \tag{5-19}$$
其中$\psi(\cdot)$是变元的q维矢值函数．

主方的性能指标为
$$J_L(u(\cdot), v(\cdot)) = F_1(x(t_f), t_f) + \int_{t_0}^{t_f} L_1(x(t), u(t), v(t), t)\mathrm{d}t, \tag{5-20}$$

其中 $F_1(\cdot\cdot), L_1(\cdot\cdot\cdot\cdot)$ 为标量函数，而从方的性能指标为

$$J_F(u(\cdot), v(\cdot)) = F_2(x(t_f), t_f) + \int_{t_0}^{t_f} L_2(x(t), u(t), v(t), t)\mathrm{d}t, \quad (5\text{-}21)$$

其中 $F_2(\cdot\cdot), L_2(\cdot\cdot\cdot\cdot)$ 为标量函数.

记容许策略集为 $\mathscr{U}_L \times \mathscr{V}_F \stackrel{\text{def}}{=} \{(u(t), v(t))\}$,
其中 $u(t), v(t)$ 为定义在$[t_0, t_f]$上的取值于 U, V 的分段连续 r, p 维矢值函数. 它们使(5-18)式的初值条件为 x_0 的解存在且唯一，并在 t_f 时刻满足(5-19)式.

$(u(t), v(t)) \in \mathscr{U}_L \times \mathscr{V}_F$　系指 $u(t) \in \mathscr{U}_L, v(t) \in \mathscr{V}_F$.

系统(5-18)式、(5-19)式和性能指标(5-20)式、(5-21)式的斯蒂克贝格最优策略问题系指：

(1) 寻找从 U 到 V 的连续映射 T，使得对于每一个 $u(t) \in \mathscr{V}_L$，成立

$$J_2(u(\cdot), T(u(\cdot))) \leqslant J_2(u(\cdot), v(\cdot)) \qquad (\forall v(t) \in \mathscr{V}_F). \qquad (5\text{-}22)$$

(2) 在 $\mathscr{V}_L$ 中寻找 $u^*(t)$，使它和 $v^*(t) = T(u^*(t))$ 一起满足

$$J_1(u^*(\cdot), v^*(\cdot)) \leqslant J_1(u(\cdot) T(u(\cdot))) \qquad (\forall u(t) \in \mathscr{V}_L). \qquad (5\text{-}23)$$

如果满足(5-22)式和(5-23)式的 T 和 $u^*(t)$ 存在，则称$(u^*(t), v^*(t)) = T(u^*(t)))$ 为斯蒂克贝格问题的最优策略，T 称为从方对主方的响应. (5-18)式对应$(u^*(t), v^*(t))$的解 $x^*(t)$ 称为最优轨线.

斯蒂克贝格策略问题虽然是一类微分对策，但由于其最优策略问题中包含一个从 U 到 V 的待求映射 T，前面给出的(单方或双方)极值原理尚不足以完全刻画其最优策略应满足的必要条件，因此，此类策略问题被认为是"非经典"的最优控制问题. 对于某些特殊的系统，在特定假设下，其斯蒂克贝格最优策略问题已有些研究结果. 但对于一般性质的问题，还无任何研究结果，这是一个值得深入研究的课题.

5.4.2　微分对策中的"协调"策略

有一类微分对策问题：局中人为甲、乙、丙三方，它们有相互联系的动态方程和对策结束时的终端条件，但它们又有各自的策略集和性能指标. 作为协调者的甲方，每给定一个策略，乙方和丙方皆给出响应，即乙方和丙方各自选择自己的策略，使甲方给定的这个策略对自己有利(使自己的性能指标达到最小). 在乙方和丙方响应的情况下，甲方再选择使自己最有利的策略. 这就是微分对策中的"协调". 这里，作为协调者的甲方可视为主方，而将乙、丙两方视为从方. 因此，微分对策中的"协调"策略问题可作为主、从微分对策问题来处理.

"协调"策略问题可用状态方程表示为

$$\begin{aligned} \dot{x} &= f(x, u, v, w, t); \\ x(t_0) &= x_0. \end{aligned} \qquad (5\text{-}24)$$

其中 $x \in \mathbf{R}^n$ 是联系甲、乙、丙三方的状态；$u \in \mathbf{R}^r, v \in \mathbf{R}^p, w \in \mathbf{R}^q$ 分别为甲、乙、丙的策略，它们分别在有界闭集 $U(\in \mathbf{R}^r), V(\in \mathbf{R}^p), W(\in \mathbf{R}^q)$ 中取值；$f(\cdot\cdot\cdot\cdot)$ 是其变元的 n 维矢值函数，并保证微分方程(5-24)式的初值解存在且唯一.

在“协调”结束时刻 t_f 时，终端状态满足的条件，即目标集为

$$\psi(x(t_f),t_f)=0_l,\tag{5-25}$$

其中 $\psi\in\mathbf{R}^l$，即 $\psi(\cdot\cdot)$ 是变元的 l 维矢值函数.

甲、乙、丙三方的性能指标分别为

$$J_1(u(\cdot),v(\cdot),w(\cdot))=F_1(x(t_f),t_f)+\int_{t_0}^{t_f}L_1(x(t),u(t),v(t),w(t),t)\mathrm{d}t,\tag{5-26}$$

$$J_2(u(\cdot),v(\cdot),w(\cdot))=F_2(x(t_f),t_f)+\int_{t_0}^{t_f}L_2(x(t),u(t),v(t),w(t),t)\mathrm{d}t,\tag{5-27}$$

$$J_3(u(\cdot),v(\cdot),w(\cdot))=F_3(x(t_f),t_f)+\int_{t_0}^{t_f}L_3(x(t),u(t),v(t),w(t),t)\mathrm{d}t.\tag{5-28}$$

记容许策略集

$$\mathscr{U}_{甲}\times\mathscr{V}_{乙}\times\mathscr{W}_{丙}\overset{\text{def}}{=\!=}\{(u(t),v(t),w(t))\},$$

其中 $u(t),v(t),w(t)$ 皆是定义在 $[t_0,t_f]$ 上的，分别从 U,V,W 中取值的 r,p,q 维分段连续函数. 它们使(5-24)式的解存在且唯一，并在 t_f 时满足(5-25)式.

$(u(t),v(t),w(t))\in\mathscr{U}_{甲}\times\mathscr{V}_{乙}\times\mathscr{W}_{丙}$ 系指：$u(t)\in\mathscr{U}_{甲}$，$v(t)\in\mathscr{U}_{乙}$，$w(t)\in\mathscr{W}_{丙}$.

系统(5-24)式，(5-25)式及指标(5-26)式 ～ (5-28)式的最优“协调”问题系指：

(1) 寻找从 U 到 V 和 W 的连续映射 T_1 和 T_2，使得对于每个 $u(t)\in\mathscr{U}_{甲}$，成立下列不等式：

$$\begin{aligned}&J_2(u(\cdot),T_1(u(\cdot)),T_2(u(\cdot)))\\&\leqslant J_2(u(\cdot),v(\cdot),T_2(V(\cdot)))\qquad(\forall v(t)\in\mathscr{U}_{乙}),\end{aligned}\tag{5-29}$$

$$\begin{aligned}&J_3(u(\cdot),T_1(u(\cdot))),T_2(u(\cdot))\leqslant J_3(u(\cdot),T_1(u(\cdot)),w(\cdot))\\&(\forall w(t)\in\mathscr{W}_{丙}).\end{aligned}\tag{5-30}$$

(2) 在 $\mathscr{U}_{甲}$ 中寻找 $u^*(t)$，使它和 $v^*(t)\overset{\text{def}}{=\!=}T_1(u^*(t)),w^*(t)\overset{\text{def}}{=\!=}T_2(u^*(t))$ 一起满足

$$\begin{aligned}&J_1(u^*(\cdot),v^*(\cdot),w^*(\cdot))\leqslant J_1(u(\cdot),T_1(u(\cdot)),T_2(u(\cdot)))\\&(\forall u\in\mathscr{U}_{甲}).\end{aligned}\tag{5-31}$$

当满足(5-29)式 ～ (5-31)式的 T_1,T_2 和 $u^*(t)$ 存在时，$(u^*(t),v^*(t),w^*(t))$ 称为“协调”问题的最优策略，相应的(5-24)式的解 $x^*(t)$ 称为最优轨线. $(x^*(t),u^*(t),v^*(t),w^*(t))$ 称为最优解.

对于最优协调问题，要研究其最优解应满足的必要条件，存在着与斯蒂克贝格最优策略问题相同的困难.

参考文献

1 Понтрягин Л С, Болтянский В Г, Гамкрелидзе Р В, Мищенко Е ф. Математигеская теория оптимальных ироцессоь. москьа: Гос. Издотельстьо физмат. литратиры, 1961.

2 (美)庞特里亚金 L C 等著. 最佳过程的数学理论. 陈祖浩, 贺建勋, 黄光远等译. 上海: 上海科技出版社, 1965.

3 Athans M, Falbp L. Optimal comtorl, New York: McGraw-Hill, 1966.

4 张嗣瀛著. 微分对策. 北京: 科学出版社, 1987.

·经济数学卷·

第 14 篇

卡尔曼滤波

编　者　王恩平
审校者　贾沛璋

目　　录

引　言

卡尔曼(R. E. Kalman)滤波方法是 20 世纪 60 年代初在现代控制理论的发展过程中产生的一种滤波方法,它有别于维纳(N. Wiener)滤波,前者可适应于非平稳过程,后者只适用于平稳过程. 卡尔曼滤波方法在导弹、航天、航空和航海等许多技术领域里都有广泛的应用. 它对促进现代控制理论的形成和发展起着举足轻重的作用,在随机控制、系统辨识和适应性控制方面也有重要的应用. 离散时间随机线性系统的卡尔曼滤波方法主要是卡尔曼本人提出的,而连续时间随机线性系统的卡尔曼滤波方法则是卡尔曼和布什(R. Bush)两人合作研究的结果,因此,也称卡尔曼滤波器为卡尔曼-布什滤波器.

1　离散时间随机线性系统的卡尔曼滤波

1.1　数学模型

1.1.1　状态方程和量测方程

离散时间线性系统的卡尔曼滤波方法适用于研究离散时间随机线性系统的状态估计. 通常一个离散时间随机线性系统的数学模型由它的状态方程和量测方程给出,即

$$\Sigma_d: \boldsymbol{x}_k = \boldsymbol{\Phi}_{k,k-1}\boldsymbol{x}_{k-1} + \boldsymbol{w}_{k-1},$$
$$\boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k.$$

其中 $\boldsymbol{x}_k$ 表示系统 Σ_d 的 n 维随机**状态向量**;$\boldsymbol{y}_k$ 表示系统 Σ_d 的 m 维随机**量测输出向量**;$\boldsymbol{\Phi}_{k,k-1}$是一个 $n\times n$ 阶矩阵,称为系统 Σ_d 的**状态转移矩阵**;$\boldsymbol{H}_k$ 是一个 $m\times n$ 阶矩阵,称为系统 Σ_d 的**量测矩阵或输出矩阵**,$\{\boldsymbol{w}_k, k=0,1,2,\cdots\}$表示加在系统 Σ_d 上的外部随机干扰序列,称为**模型噪声**序列,对于任意 $k\geqslant 0$,$\boldsymbol{w}_k$ 是一个 n 维随机矢量;$\{\boldsymbol{v}_k, k=1,2,\cdots\}$表示系统 Σ_d 的随机量测误差序列,称为**量测噪声**序列,$\boldsymbol{v}_k$ 是一个 m 维随机向量. $\boldsymbol{w}_k$ 和 $\boldsymbol{v}_k$ 分别代表 t_k 时刻的模型噪声向量和量测输出噪声向量.

通常称描述系统 Σ_d 的第一个差分方程为系统的**状态方程**,第二个代数方程为系统的量测输出方程,或简称**量测方程**.

1.1.2　状态转移矩阵

状态转移矩阵是刻画两个不同时刻系统状态之间关联关系的矩阵,对系统 Σ_d

而言，$\Phi_{k,k-1}$就是状态转移矩阵，它刻画了系统 Σ_d 的状态是怎样从 t_{k-1}时刻演化到 t_k 时刻的. 一个线性系统的状态转移矩阵 $\Phi_{k,k-1}$有如下三个性质：

1° 传递性　即对于任意 $k>j>i\geqslant 0$，都有

$$\Phi_{k,i}=\Phi_{k,j}\Phi_{j,i}.$$

2° 可逆性　即对于任意 $k>j\geqslant 0$，都有

$$\Phi_{k,j}^{-1}=\Phi_{j,k},$$

其中"－1"表示矩阵的逆.

3° 对于任意 $k\geqslant 0$，都有 $\Phi_{k,k}=I$ 为 $n\times n$ 阶单位矩阵.

综合状态转移矩阵的上述性质，可以得到

$$\Phi_{k,i}=\Phi_{k,j}\Phi_{j,i}\qquad(\forall\, k,j,i).$$

必须指出，状态转移矩阵的可逆性通常是自然的，因为一般来说，离散时间线性系统都是从连续时间线性系统经离散化处理后得到的. 但有时依客观对象的实际情况也可以直接建立离散时间线性系统的状态方程，这时其状态转移矩阵可能不可逆，本篇一般不讨论这种情况.

状态转移矩阵的上述三个性质在研究卡尔曼滤波方法中有着重要作用.

1.1.3 基本假设

在推导卡尔曼滤波方法之前，对数学模型需要做一些必要的基本假设，以便简化问题的讨论. 这些基本假设为：

(1)假设系统 Σ_d 的初始状态向量 x_0 是一个服从正态分布的随机向量，其均值和协方差有穷，即

$$E\{x_0\}=\bar{x}_0,$$

$$E\{x_0x_0^{\mathrm{T}}\}=P_0.$$

其中 $E\{\cdot\}$表示期望算子；P_0 是一个 $n\times n$ 阶对称非负定矩阵；"T 表示向量或矩阵的转置.

(2)假设系统 Σ_d 的模型噪声$\{w_k,k=0,1,2,\cdots\}$是一个均值为零、协方差有穷的独立正态随机序列，或称均值为零的白噪声序列. 即对于任意 $k\geqslant j\geqslant 0$，有

$$E\{w_k\}=0,$$

$$E\{w_k,w_j^{\mathrm{T}}\}=Q_k\delta_{k,j}.$$

其中对于任意 $k\geqslant 0$，Q_k 都是一个 $n\times n$ 阶对称非负定矩阵；

$$\delta_{k,j}=\begin{cases}1 & (k=j);\\ 0 & (k\neq j).\end{cases}$$

(3)假设系统 Σ_d 的量测噪声$\{v_k,k=1,2,\cdots\}$是一个均值为零、协方差有穷的独立正态随机序列. 或称均值为零的白噪声序列. 即对于任意 $k\geqslant j\geqslant 1$，有

$$E\{v_k\}=0,$$

$$E\{v_k,v_j^{\mathrm{T}}\}=R_k\delta_{k,j}.$$

其中对于任意 $k\geqslant 1$，R_k 都是一个 $m\times m$ 阶对称正定矩阵，因而非奇异.

(4)假设系统 Σ_d 的初始状态与模型噪声和量测噪声都是相互独立的，模型噪

声和量测噪声之间也是相互独立的,即对于任意 $k,j\geqslant 0$,都有

$$E\{x_0w_k^{\mathrm{T}}\}=0,$$
$$E\{x_0v_k^{\mathrm{T}}\}=0,$$
$$E\{w_kv_j^{\mathrm{T}}\}=0.$$

下面称满足上述四项假设的系统 Σ_{d} 为正态系统或高斯(G.F.Gauss)系统.

1.2 系统状态的最优估计

1.2.1 问题的叙述

所谓状态估计问题是指,给出量测数据序列$\{y_j,j=1,2,\cdots,k\}$后,求系统 Σ_{d} 的状态向量 x_i 在某种意义下的最优估计,记作 $\hat{x}_{i|k}$.如果 $i>k$,则 $\hat{x}_{i|k}$叫做x_i 的**预测估计**;如果 $i=k$,则 $\hat{x}_{k|k}$叫做x_k 的**滤波估计**;如果 $i<k$,则 $\hat{x}_{i|k}$叫做x_i 的**平滑估计**.如果 $\hat{x}_{i|k}$使得

$$E\{[x_i-\hat{x}_i]^{\mathrm{T}}[x_i-\hat{x}_i]\}$$

达到极小,那么 $\hat{x}_{i|k}$叫做x_i 的**最小方差估计**,或称**最优估计**,其中 $\hat{x}_i$ 代表任意 n 维随机向量.

如果 $\hat{x}_{i|k}$是x_i 的最优估计,那么定义

$$P_{i,k}=E\{[x_i-\hat{x}_{i|k}][x_i-\hat{x}_{i|k}]^{\mathrm{T}}\}$$

为估计误差 $\tilde{x}_{i|k}=x_i-\hat{x}_{i|k}$的协方差矩阵,理论上它是估计精度的一个度量.

如果 $E\{x_i\}=E\{\hat{x}_{i|k}\}$,那么 $\hat{x}_{i|k}$叫做x_i 的**无偏估计**.

如果 $\hat{x}_{i|k}$是量测数据$y_1,y_2,\cdots,y_k$ 的线性函数,那么 $\hat{x}_{i|k}$叫做x_i 的**线性估计**.

卡尔曼滤波方法所要解决的问题是,给出量测数据序列$\{y_j,j=1,2,\cdots,k\}$后,求系统 Σ_{d} 状态向量 x_k 的最优估计$\hat{x}_{k|k}$.

1.2.2 最优估计与条件期望

依据统计学原理可知,给定量测数据序列$\{y_j,j=1,2,\cdots,k\}$后,系统 Σ_{d} 的状态向量 x_i 的最优估计为

$$\hat{x}_{i|k}=E\{x_i|y_1,y_2,\cdots,y_k\},$$

其中 $E\{x_i|y_1,y_2,\cdots,y_k\}$表示 x_i 关于$y_1,y_2,\cdots,y_k$ 的**条件期望**.

当 $i=k$ 时,$\hat{x}_{k|k}=E\{x_k|y_1,y_2,\cdots,y_k\}$就是系统 Σ_{d} 状态向量 x_k 的最优滤波估计,当 $i=k+1$ 时,$E\{x_{k+1}|y_1,y_2,\cdots,y_k\}$就是系统 Σ_{d} 状态向量 x_{k+1}的一步最优预测估计.

假设系统 Σ_d 是正态的,则条件期望 $E\{x_i|y_1,y_2,\cdots,y_k\}$ 是量测数据 $y_1,y_2,\cdots,y_k$ 的线性函数.由此可见,正态系统状态的最优估计一定是线性估计.这时就不必再刻意寻求非线性估计了,因为它绝不是最优估计.反之,如果系统 Σ_d 的状态最优估计是线性估计,那么 Σ_d 必是正态系统.因此可以说系统 Σ_d 状态的最优估计的线性性与它的正态假设是等价的.

1.3 卡尔曼滤波器

1.3.1 一步最优预测估计

给出量测数据 $y_1,y_2,\cdots,y_{k-1}$后,系统状态 x_k 的一步最优预测估计为

$$\hat{x}_{k|k-1}=E\{x_k|y_1,y_2,\cdots,y_{k-1}\},$$

由系统 Σ_d 的状态方程可知,

$$\begin{aligned}\hat{x}_{k|k-1}&=E\{\Phi_{k,k-1}x_{k-1}+w_{k-1}|y_1,y_2,\cdots,y_{k-1}\}\\&=\Phi_{k,k-1}E\{x_{k-1}|y_1,y_2,\cdots,y_{k-1}\}+\\&\quad E\{w_{k-1}|y_1,y_2,\cdots,y_{k-1}\}.\end{aligned}$$

而

$$E\{x_{k-1}|y_1,y_2,\cdots,y_{k-1}\}=\hat{x}_{k-1|k-1}$$

为系统 Σ_d 状态向量 x_{k-1}的最优滤波估计,并依基本假设,有

$$E\{w_{k-1}|y_1,y_2,\cdots,y_{k-1}\}=0,$$

于是有

$$\hat{x}_{k|k-1}=\Phi_{k,k-1}\hat{x}_{k-1|k-1}.\tag{1-1}$$

如果状态向量 x_{k-1}的最优滤波估计是无偏的,那么状态向量 x_k 的一步最优预测估计也是无偏的.

1.3.2 新息过程

如果已经得到系统 Σ_d 的状态向量 x_k 的一步最优预测估计 $\hat{x}_{k|k-1}$,在 t_k 时刻又未获得新的量测数据 y_k,则量测数据 y_k 的一步最优预测估计为

$$\begin{aligned}\hat{y}_{k|k-1}&=E\{y_k|y_1,y_2,\cdots,y_{k-1}\}\\&=E\{H_kx_k+v_k|y_1,y_2,\cdots,y_{k-1}\}\\&=H_kE\{x_k|y_1,y_2,\cdots,y_{k-1}\}+\\&\quad E\{v_k|y_1,y_2,\cdots,y_{k-1}\}.\end{aligned}$$

显然有

$$\hat{y}_{k|k-1}=H_k\hat{x}_{k|k-1}.\tag{1-2}$$

当 t_k 时刻获得实际量测数据 y_k 后,它与一步最优预测估计 $\hat{y}_{k|k-1}$之差

$$\tilde{y}_{k|k-1} = y_k - \hat{y}_{k|k-1} = y_k - H_k\hat{x}_{k|k-1}$$

构成的时间序列$\{\tilde{y}_{k|k-1}, k = 1,2,\cdots\}$被称为**新息过程**. 可以证明,新息过程是一个均值为零的白噪声序列.

1.3.3 最优滤波估计

若给出量测数据 $y_1, y_2, \cdots, y_k$,则系统 Σ_d 状态向量 x_k 的最优滤波估计为

$$\hat{x}_{k|k} = E\{x_k | y_1, y_2, \cdots, y_k\}.$$

依假设,系统 Σ_d 是正态的,再根据条件期望的性质,有

$$\begin{aligned} & E\{x_k | y_1, y_2, \cdots, y_k\} \\ & = E\{x_k | y_1, y_2, \cdots, y_{k-1}\} + E\{x_k | \tilde{y}_{k|k-1}\} - E\{x_k\}. \end{aligned}$$

于是,有

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k[y_k - H_k\hat{x}_{k|k-1}]. \tag{1-3}$$

这就是系统 Σ_d 的状态向量 x_k 的最优滤波估计所满足的递推方程,通常称 K_k 为最优增益矩阵,它要由估计 $\hat{x}_{k|k}$的最优性来确定.

1.3.4 最优增益矩阵

定义最优滤波估计误差为

$$\tilde{x}_{k|k} = x_k - \hat{x}_{k|k}.$$

经计算,可得

$$\tilde{x}_{k|k} = [I - K_kH_k]\ \tilde{x}_{k|k-1} + K_kV_k.$$

其中

$$\tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1}.$$

再定义

$$P_{k,k} = E\{\tilde{x}_{k|k}\tilde{x}_{k|k}^{\mathrm{T}}\},$$

$$P_{k,k-1} = E\{\tilde{x}_{k|k-1}\tilde{x}_{k|k-1}^{\mathrm{T}}\},$$

分别为最优滤波估计误差和一步最优预测估计误差的协方差矩阵. 经计算,可得

$$P_{k,k} = [I - K_kH_k]\ P_{k,k-1}[I - K_kH_k]^{\mathrm{T}} + K_kR_kK_k^{\mathrm{T}}.$$

用配方法可求得使 $P_{k,k}$达到极小的增益矩阵为

$$K_k = P_{k,k-1}H_k^{\mathrm{T}}[H_kP_{k,k-1}H_k^{\mathrm{T}} + R_k]^{-1}. \tag{1-4}$$

这就是递推方程(1-3)式中的最优增益矩阵所满足的代数方程.

经计算,最后可得

$$P_{k,k-1} = \Phi_{k,k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}} + Q_{k-1}, \tag{1-5}$$

$$P_{k,k} = [I - K_kH_k]\ P_{k,k-1}. \tag{1-6}$$

1.3.5 卡尔曼滤波器

(1-1)式~(1-6)式,通常被称为系统 Σ_d 的卡尔曼滤波算法,它可给出系统状态向量的一步最优预测估计和最优滤波估计.由(1-1)式和(1-3)式,有

$$\hat{x}_{k|k}=\Phi_{k,k-1}\hat{x}_{k-1|k-1}+K_k[y_k-H_k\Phi_{k,k-1}\hat{x}_{k-1|k-1}], \tag{1-7}$$

或者

$$\hat{x}_{k|k}=[I-K_kH_k]\ \Phi_{k,k-1}\hat{x}_{k-1|k-1}+K_ky_k. \tag{1-8}$$

由此可见,上式给出的递推公式同时描述了一个离散时间动态系统,所以,有时也把卡尔曼滤波算法称为卡尔曼滤波器,它的动力学方程为

$$\begin{aligned}
&\hat{x}_{k|k}=[I-K_kH_k]\ \Phi_{k,k-1}\hat{x}_{k-1|k-1}+K_ky_k,\\
&K_k=P_{k,k-1}H_k^{\mathrm{T}}[H_kP_{k,k-1}H_k^{\mathrm{T}}+R_k]^{-1},\\
&P_{k,k-1}=\Phi_{k,k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}}+Q_{k-1},\\
&P_{k,k}=[I-K_kH_k]\ P_{k,k-1};
\end{aligned}$$

初始条件为

$$\begin{aligned}
&\hat{x}_{0|0}=E\{x_0\}=\bar{x}_0,\\
&P_{0,0}=E\{x_0x_0^{\mathrm{T}}\}=P_0.
\end{aligned}$$

对于卡尔曼滤波器,在应用时需注意如下几点:

(1)由卡尔曼滤波器给出的状态估计是最小方差估计.

(2)在研究离散时间随机线性系统的卡尔曼滤波方法时,假设 Σ_d 是正态系统,因此所给出的系统状态 x_k 的最优估计是量测数据的线性函数,即最优估计是线性估计.如果对 Σ_d 取消正态系统的假设,那么由(1-3)式给出的系统状态 x_k 的估计 $\hat{x}_{k|k}$是线性最小方差估计,也就是说,它是在线性估计的范围内是最优的.

(3)卡尔曼滤波器是一个**无限记忆滤波器**,即历史的量测数据对未来的状态估计总是起作用的.

(4)卡尔曼滤波器的增益矩阵还有另一种形式,即

$$K_k=P_{k,k}H_k^{\mathrm{T}}R_k^{-1}.$$

(5)由卡尔曼滤波器给出的状态向量 x_k 的线性最小方差估计 $\hat{x}_{k|k}$有明显的几何意义.事实上,如果用量测数据 $y_1,y_2,\cdots,y_k$ 组成一个 n 维线性子空间

$$z_k=\{z_k:z_k=\sum_{i=1}^{k}B_{k,i}\quad y_i\},$$

其中$B_{k,i}$是任意 $n\times m$ 阶矩阵.于是,可以证明 $\hat{x}_{k|k}$是 x_k 在线性子空间 z_k 上的**正交投影**,记为

$$\hat{x}_{k|k}=\hat{E}\{x_k|z_k\}.$$

显然有

$$z_{k-1}\subset z_k.$$

由于 z_k 是一个希尔伯特(D. Hilbert)空间,因此,依正交投影定理,有

$$z_k = z_{k-1} \oplus \tilde{y}_k.$$

其中,$\tilde{y}_k$ 是 z_{k-1} 的正交补空间,它的每一个元都具有 $\boldsymbol{K}_k[y_k - \boldsymbol{H}_k\hat{\boldsymbol{x}}_{k|k-1}]$ 的形式,于是有

$$\hat{\boldsymbol{x}}_{k|k} = \hat{E}\{x_k | z_{k-1}\} + \hat{E}\{\hat{x}_k | \tilde{y}_k\},$$

同样可得出

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k[y_k - \boldsymbol{H}_k\hat{\boldsymbol{x}}_{k|k-1}].$$

其中 $\boldsymbol{K}_k$ 就是待定的增益矩阵,它同样可以按前面列出的公式计算.

2 卡尔曼滤波器的稳定性

2.1 一致完全能控性和一致完全能观测性

定义 1 假定离散时间随机线性系统为 Σ_d,满足 1.1.3 小节的假设.如果存在正常数 α_1 和 β_1 以及正整数 N,使得对于一切 $k \geqslant N$,都有

$$\alpha_1 \boldsymbol{I} \leqslant \sum_{i=k-N}^{k-1} \boldsymbol{\Phi}_{k,i+1}\boldsymbol{Q}_{i+1}\boldsymbol{\Phi}_{k,i+1}^{\mathrm{T}} \leqslant \beta_1 \boldsymbol{I},$$

其中 $\boldsymbol{I}$ 表示具有相应阶数的单位矩阵,那么称系统 Σ_d 是**一致完全能控**的.

定义 2 假定离散时间随机线性系统为 Σ_d,满足 1.1.3 小节的假设.如果存在正常数 α_2 和 β_2 以及正整数 N,使得对于一切 $k \geqslant N$,都有

$$\alpha_2 \boldsymbol{I} \leqslant \sum_{i=k-N}^{k} \boldsymbol{\Phi}_{i,k}^{\mathrm{T}}\boldsymbol{H}_i^{\mathrm{T}}\boldsymbol{R}_i^{-1}\boldsymbol{H}_i\boldsymbol{\Phi}_{i,k} \leqslant \beta_2 \boldsymbol{I},$$

那么称系统 Σ_d 是**一致完全能观测**的.

定义 3 假设离散时间随机线性系统为 Σ_d,如果对于任意 $k \geqslant 0$,都有

$$\boldsymbol{\Phi}_{k,k-1} = \boldsymbol{\Phi},\ \boldsymbol{H}_k = \boldsymbol{H}$$

为常值矩阵,且 $\boldsymbol{\Phi}$ 可逆,则称它为**定常系统**.如果对于任意 $k \geqslant 0$,都有

$$\boldsymbol{Q}_k = \boldsymbol{Q} \geqslant 0,\ \boldsymbol{R}_k = \boldsymbol{R} > 0$$

为常值矩阵,则称它为**平稳系统**.如果 Σ_d 既是定常的又是平稳的,则称它为平稳定常系统.

对于平稳定常系统而言,只要$[\boldsymbol{\Phi}, \boldsymbol{Q}]$**完全能控**,它一定一致完全能控;只要$[\boldsymbol{\Phi}, \boldsymbol{H}]$**完全能观测**,它一定一致完全能观测.

需要指出,在定义 3 中为使系统是定常的,未必要求 $\boldsymbol{\Phi}$ 可逆,但是如果 $\boldsymbol{\Phi}$ 不可逆,那么即使$(\boldsymbol{\Phi}, \boldsymbol{Q})$完全能控.也不能保证系统 Σ_d 一致完全能控.

2.2 稳定性定理

定理1 假定离散时间随机线性系统为 Σ_d,满足1.1.3小节的假设.如果它是一致完全能控和一致完全能观测的,那么系统 Σ_d 的卡尔曼滤波器(1-1)式~(1-8)式有如下性质:

1° 对于任意 t_0 时刻的初始协方差矩阵 $\boldsymbol{P}_0 \geqslant 0$,矩阵里卡蒂(J.F.Riccati)方程(1-4)式~(1-6)式都有唯一对称非负定解 $\boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0)$,而且当 $k>N$ 时,$\boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0)>0$.这里 $\boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0)$ 表示以 t_0 为初始时刻,$\boldsymbol{P}_0$ 为初始条件由(1-4)式~(1-6)式迭代而得到的解.

2° 对于任意 t_0 时刻的初始协方差矩阵 $\boldsymbol{P}_0 \geqslant 0$,当 $k>N$ 时,$\boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0)$ 一致有界,即存在正常数 α_3 和 β_3,使得对于一切 $k \geqslant N$,都有

$$\alpha_3 \boldsymbol{I} \leqslant \boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0) \leqslant \beta_3 \boldsymbol{I}.$$

3° 对于任意 t_0 时刻的初始协方差矩阵 $\boldsymbol{P}_0 \geqslant 0$,都有

$$\lim_{t_0 \to -\infty} \boldsymbol{P}_{k,k}(t_0;\boldsymbol{P}_0) = \widetilde{\boldsymbol{P}}_{k,k} > 0.$$

4° 卡尔曼滤波器(1-8)式是大范围一致渐近稳定的(或按指数衰减稳定的),即若令

$$\boldsymbol{\Psi}_{k,k-1} = [\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k]\ \boldsymbol{\Phi}_{k,k-1}$$

为状态转移矩阵,且

$$\boldsymbol{\Psi}_{k,j} = \boldsymbol{\Psi}_{k,k-1} \cdot \boldsymbol{\Psi}_{k-1,k-2} \cdots \boldsymbol{\Psi}_{j+1,j},$$

则存在正常数 α_4 和 β_4,使得对于一切 $k>j$,都有

$$\| \boldsymbol{\Psi}_{k,j} \| \leqslant \alpha_4 \exp(-\beta_4(t_k - t_j)),$$

其中 $\|\cdot\|$ 表示矩阵的算子范数.

注1 在定理1中的 N,就是使得系统 Σ_d 同时保证一致完全能控和一致完全能观测的那个正整数.

注2 定理1的性质1°表明,从理论上讲,只要系统 Σ_d 存在模型噪声和量测噪声,那么由卡尔曼滤波器给出的状态估计不会完全精确,必含有估计误差.再结合其性质2°可知,其估计精度也不能太差.性质3°和性质4°表明了卡尔曼滤波器对初始条件的依赖关系.

对于平稳定常系统,定理1可改写为下面的定理2.

定理2 假设离散时间随机线性系统 Σ_d 是平稳的和定常的.如果 $[\boldsymbol{\Phi},\boldsymbol{Q}]$ 完全能控,$[\boldsymbol{\Phi},\boldsymbol{H}]$ 完全能观测,那么里卡蒂方程

$$\boldsymbol{P} = [\boldsymbol{I} - \boldsymbol{P}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{H}][\boldsymbol{\Phi}\boldsymbol{P}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{Q}], \tag{2-1}$$

有唯一对称正定解 $\boldsymbol{P}$,如果令

$$\boldsymbol{K} = \boldsymbol{P}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1} \tag{2-2}$$

作为增益矩阵,则稳态卡尔曼滤波器

$$\widehat{\boldsymbol{x}}_{k|k} = [\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H}]\ \boldsymbol{\Phi}\,\widehat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}\,\boldsymbol{y}_k \tag{2-3}$$

是稳定的,即矩阵$[\boldsymbol{I}-\boldsymbol{KH}]\boldsymbol{\Phi}$的特征值都在复平面的单位圆内.

注 3 稳态卡尔曼滤波器也称为**常增益卡尔曼滤波器**,由(2-2)式给出的矩阵$\boldsymbol{K}$称为卡尔曼滤波器的**稳态增益**.其实,稳态卡尔曼滤波器也可看做是最优的观测器.在应用上采用常增益的卡尔曼滤波器是很方便的,它可以离线计算最优增益矩阵,而无须迭代运算,从而大大地减少计算量.为了说明它的优越性,下面列举两个常用的最简单的卡尔曼滤波器的例子来说明.

1. α-β 滤波器

在许多场合下,一个常用的最简单的滤波器就是 **α-β 滤波器**,它可以解决能够用一次曲线逼近的信号处理问题.例如,雷达指挥仪测得的匀速直线飞行的飞机位置数据处理问题就可以采用这种滤波器.

假设系统的状态方程和量测方程分别为

$$\boldsymbol{x}_k=\boldsymbol{\Phi}\,\boldsymbol{x}_{k-1}+\boldsymbol{b}\,w_{k-1},\tag{2-4}$$

$$y_k=\boldsymbol{H}\,\boldsymbol{x}_k+v_k.\tag{2-5}$$

其中$\boldsymbol{x}_k$为二维状态向量;y_k为一标量量测输出;$\{w_k,k=0,1,2,\cdots\}$为均值为零、方差为σ^2的正态随机序列;$\{v_k,k=1,2,\cdots\}$均值为零、方差为r^2的正态随机序列;同时

$$\boldsymbol{\Phi}=\begin{bmatrix}1&T\\0&1\end{bmatrix};\quad \boldsymbol{b}=\begin{bmatrix}0\\1\end{bmatrix};\quad \boldsymbol{H}=[1,0].$$

T为采样时间间隔.假设对于任意$k,j\geqslant 0$,都有

$$E\{w_kv_j\}=0,\ E\{x_0w_k\}=0,\ E\{x_0v_j\}=0.$$

于是这个最简单的二阶系统的卡尔曼滤波器方程为

$$\hat{\boldsymbol{x}}_{k|k}=[\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H}]\,\boldsymbol{\Phi}\,\hat{\boldsymbol{x}}_{k-1|k-1}+\boldsymbol{K}_ky_k.\tag{2-6}$$

这里,最优增益矩阵$\boldsymbol{K}_k$只有两个可变参数,记为

$$\boldsymbol{K}_k=\begin{bmatrix}\alpha_k\\\beta_k\end{bmatrix}.$$

按照卡尔曼滤波递推公式,可以算出

$$\alpha_k=\frac{p_{k,k-1}(1,1)}{p_{k,k-1}(1,1)+r^2},\tag{2-7}$$

$$\beta_k=\frac{p_{k,k-1}(1,2)}{p_{k,k-1}(1,2)+r^2},\tag{2-8}$$

其中$p_{k,k-1}(i,j)$表示矩阵$\boldsymbol{P}_{k,k-1}$中的第i行第j列上的元素.注意到

$$\boldsymbol{K}_k=\boldsymbol{P}_{k,k}\boldsymbol{H}^{\mathrm{T}}r^{-2},$$

所以有

$$\alpha_k=\frac{p_{k,k}(1,1)}{r^2},$$

$$\beta_k=\frac{p_{k,k}(1,2)}{r^2}.$$

由于$[\boldsymbol{\Phi},\boldsymbol{b}]$能控,$[\boldsymbol{\Phi},\boldsymbol{H}]$能观测,因此依定理2可知,$\alpha_k$ 和β_k 有稳态解,分别为

$$\alpha=\frac{p_{11}}{r^2},$$

$$\beta=\frac{p_{12}}{r^2}.$$

其中

$$p_{11}=\lim_{k\to\infty}p_{k,k}(1,1),$$

$$p_{12}=\lim_{k\to\infty}p_{k,k}(1,2).$$

或者

$$\alpha=\frac{p'_{11}}{p'_{11}+r^2},$$

$$\beta=\frac{p'_{12}}{p'_{12}+r^2}.$$

其中 $p'_{11}=\lim\limits_{k\to\infty}p_{k,k-1}(1,1)$;$p'_{12}=\lim\limits_{k\to\infty}p_{k,k-1}(1,2)$.

这时稳态卡尔曼滤波器方程可改写为

$$\hat{x}^1_{k|k}=(1-\alpha)\hat{x}^1_{k-1|k-1}+(1-\alpha)T\hat{x}^2_{k-1|k-1}+\alpha y_k,$$

$$\hat{x}^2_{k|k}=-\beta\hat{x}^1_{k-1|k-1}+(1-\beta T)\hat{x}^2_{k-1|k-1}+\beta y_k.$$

由于这个滤波器只依赖于两个增益常数 α 和β,故得名 α-β 滤波器.迭代计算(2-7)式和(2-8)式,最后可得出 α 和β 的稳态值.

还须指出,α-β 滤波器的稳态增益实际上只取决于信噪比 $h=\frac{\sigma}{r}$,而与 σ 和 r 的具体取值无关.这是不难看出的,只要取 $\boldsymbol{P}_0=\sigma^2\boldsymbol{I}$,在卡尔曼滤波的递推算法中用归纳法可以说明,$\boldsymbol{K}$ 仅仅是 h^2 的函数,因此 α 和β 只依赖于 h,滤波器的品质也仅仅依赖于信噪比 h.

2. α-β-γ 滤波器

α-β 滤波器适用于那些能够用一次曲线逼近的信号处理问题,但是在许多实际应用中,一阶近似不能很好地刻画一批量测数据的变化趋势,而需要用二次曲线或更高次曲线来逼近.**α-β-γ 滤波器**就是一种能够用于二次曲线逼近的数据处理方法.

假设系统的状态方程和量测方程仍由(2-4)式~(2-5)式表示,不过这时 $\boldsymbol{x}_k$ 是一个三维状态向量,同时

$$\boldsymbol{\Phi}=\begin{bmatrix}1 & T & \frac{T^2}{2}\\ 0 & 1 & T\\ 0 & 0 & 1\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}0\\0\\1\end{bmatrix},\quad \boldsymbol{H}=[1\ 0\ 0].$$

于是系统(2-4)式、(2-5)式的卡尔曼滤波器方程仍具有(2-6)式那种表达形式,只是最优增益矩阵 $\boldsymbol{K}_k$ 依赖于三个参数,记作

$$K_k=\begin{bmatrix}\alpha_k\\ \beta_k\\ \gamma_k\end{bmatrix}.$$

同样可以算出：

$$\alpha_k=\frac{p_{k,k-1}(1,1)}{p_{k,k-1}(1,1)+r^2}=\frac{p_{k,k}(1,1)}{r^2},$$

$$\beta_k=\frac{p_{k,k-1}(1,2)}{p_{k,k-1}(1,2)+r^2}=\frac{p_{k,k}(1,2)}{r^2},$$

$$\gamma_k=\frac{p_{k,k-1}(1,3)}{p_{k,k-1}(1,3)+r^2}=\frac{p_{k,k}(1,3)}{r^2}.$$

由于$[\boldsymbol{\Phi},\boldsymbol{b}]$完全能控，$[\boldsymbol{\Phi},\boldsymbol{H}]$完全能观测，因此依定理2可知，$\alpha_k,\beta_k$和$\gamma_k$有稳态值分别记作$\alpha$、$\beta$和$\gamma$，这时稳态卡尔曼滤波器方程为

$$\hat{\boldsymbol{x}}_{k|k}=\boldsymbol{\Psi}\hat{\boldsymbol{x}}_{k-1|k-1}+\boldsymbol{K}y_k,$$

其中

$$\boldsymbol{\Psi}=\begin{bmatrix}(1-\alpha) & (1-\alpha)T & (1-\alpha)\frac{T^2}{2}\\ -\beta & 1-\beta T & T(1-\frac{\beta T}{2})\\ -\gamma & -\gamma T & 1-\frac{\gamma T^2}{2}\end{bmatrix},\quad \boldsymbol{K}=\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix}.$$

3　卡尔曼滤波方法的推广

3.1　带有相关量测噪声的卡尔曼滤波方法

已知系统Σ_d，假设量测噪声$\{\boldsymbol{v}_k\}$是一个相关噪声序列，具有一阶马尔可夫(A. A. Markov)性质，可用下列随机差分方程描述：

$$\boldsymbol{v}_k=\boldsymbol{\Psi}_{k,k-1}\boldsymbol{v}_{k-1}+\boldsymbol{\eta}_{k-1},\tag{3-1}$$

其中$\boldsymbol{\Psi}_{k,k-1}$为一个$m\times m$阶矩阵；$\{\boldsymbol{\eta}_k,k=0,1,2,\cdots\}$为均值为零、协方差为$\boldsymbol{E}\{\boldsymbol{\eta}_k\boldsymbol{\eta}_j^{\mathrm{T}}\}=\boldsymbol{R}_k\delta_{k,j}$的正态白噪声序列，这里$\boldsymbol{R}_k$为$m\times m$阶对称正定矩阵. 同时假设$\boldsymbol{\eta}_k$，$\boldsymbol{w}_k$和$\boldsymbol{x}_0$都是相互独立的. 除此之外，在1.1.3小节对系统Σ_d所做的假设均不变. 现在的问题仍然是给定量测数据$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_k$后确定$\boldsymbol{x}_k$的最小方差线性无偏滤波估计. 在这种情况下，通常采用量测求差法解决这个问题. 为此令

$$\boldsymbol{z}_k=\boldsymbol{y}_{k+1}-\boldsymbol{\Psi}_{k+1,k}\boldsymbol{y}_k\tag{3-2}$$

作为新的量测数据. 将(3-1)式代入(3-2)式，有

$$\boldsymbol{z}_k=\boldsymbol{M}_k\boldsymbol{x}_k+\boldsymbol{\xi}_k,\tag{3-3}$$

其中

$$M_k = H_{k+1}\Phi_{k+1,k} - \Psi_{k+1,k}H_k,$$
$$\xi_k = \eta_k + H_{k+1}w_k.$$

可以证明，$\{\xi_k, k=1,2,\cdots\}$是一个均值为零的独立白噪声序列，它的协方差为

$$E\{\xi_k\xi_j^{\mathrm{T}}\} = (H_{k+1}Q_kH_{k+1}^{\mathrm{T}} + R_k)\delta_{k,j} \xlongequal{\text{def}} R_k^*\delta_{k,j}.$$

值得注意的是，这时量测噪声 ξ_k 和模型噪声 w_{k-1}不是相互独立的，它们之间的互协方差为

$$E\{w_k\xi_j^{\mathrm{T}}\} = Q_kH_{k+1}^{\mathrm{T}}\delta_{k,j} \xlongequal{\text{def}} C_k\delta_{k,j}.$$

这时系统

$$x_k = \Phi_{k,k-1}x_{k-1} + w_{k-1},$$
$$z_k = M_kx_k + \xi_k$$

的卡尔曼滤波器为

$$\hat{x}_{k|k} = \Phi_{k,k-1}\hat{x}_{k-1|k-1} + K_k[y_k - \Psi_{k,k-1}y_{k-1} - M_{k-1}\hat{x}_{k-1|k-1}],$$
$$K_k = [\Phi_{k,k-1}P_{k-1,k-1}M_{k-1}^{\mathrm{T}} + C_{k-1}]\cdot[M_{k-1}P_{k-1,k-1}M_{k-1}^{\mathrm{T}} + R_{k-1}^*]^{-1},$$
$$P_{k,k} = \Phi_{k,k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}} + Q_{k-1} - K_{k-1}[M_{k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}} + C_{k-1}^{\mathrm{T}}].$$

为了进行递推计算，其初始条件分别为

$$\hat{x}_{0|0} = E\{x_0\} + \overline{P}_0\overline{H}_0^{\mathrm{T}}(H_0\overline{P}_0H_0^{\mathrm{T}} + R_0)^{-1}(y_0 - H_0E\{x_0\}),$$
$$P_{0,0} = \overline{P}_0 - \overline{P}_0H_0^{\mathrm{T}}[H_0\overline{P}_0H_0^{\mathrm{T}} + R_0]^{-1}H_0\overline{P}_0,$$
$$\overline{p}_0 = E\{[x_0 - E\{x_0\}][x_0 - E\{x_0\}]^{\mathrm{T}}\}.$$

3.2 带有相关模型噪声的卡尔曼滤波方法

假设系统 Σ_d 的模型噪声序列$\{w_k, k=0,1,2,\cdots\}$是一阶马尔可夫过程，即

$$w_k = D_{k,k-1}w_{k-1} + \xi_{k-1}, \tag{3-4}$$

其中$D_{k,k-1}$为 $n\times n$ 阶矩阵；$\{\xi_k, k=0,1,2,\cdots\}$为均值为零、协方差为 $E\{\xi_k\xi_j^{\mathrm{T}}\} = Q_k\delta_{k,j}$的正态白噪声序列，$Q_k$ 为 $n\times n$ 阶对称非负定矩阵，并且对于一切 $k, j\geqslant 0$，有

$$E\{x_0\xi_k^{\mathrm{T}}\} = 0 \quad 和 \quad E\{v_k\xi_j^{\mathrm{T}}\} = 0.$$

除此之外，1.1.3 小节的假设均成立，现在的问题是给定量测数据 $y_1, y_2, \cdots, y_k$ 后确定状态向量 x_k 的最小方差线性无偏滤波估计. 为解决这个问题，可采用扩充状态变量法. 令

$$z_k = \begin{bmatrix} x_k \\ w_k \end{bmatrix}, \quad \Psi_{k,k-1} = \begin{bmatrix} \Phi_{k,k-1} & I \\ 0 & D_{k,k-1} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad M_k = [H_k, 0].$$

于是扩充系统的状态方程和量测方程分别为

$$z_k = \Psi_{k,k-1}z_{k-1} + \Gamma\xi_{k-1}, \tag{3-5}$$

$$y_k = M_k z_k + v_k. \tag{3-6}$$

将前面的结果用于扩充系统(3-5)式、(3-6)式,然后利用分块矩阵运算的性质,得到如下滤波器方程:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k^1[y_k - H_k\hat{x}_{k|k-1}],$$
$$\hat{w}_{k|k} = \hat{w}_{k|k-1} + K_k^2[y_k - H_k\hat{x}_{k|k-1}],$$
$$\hat{x}_{k|k-1} = \Phi_{k,k-1}\hat{x}_{k-1|k-1} + \hat{w}_{k-1|k-1},$$
$$\hat{w}_{k|k-1} = D_{k,k-1}\hat{w}_{k-1|k-1},$$
$$K_k^1 = P_{k,k-1}H_k^{\mathrm{T}}[H_kP_{k,k-1}H_k^{\mathrm{T}} + R_k]^{-1},$$
$$K_k^2 = G_{k,k-1}^{\mathrm{T}}H_k^{\mathrm{T}}[H_kP_{k,k-1}H_k^{\mathrm{T}} + R_k]^{-1},$$
$$P_{k,k-1} = \Phi_{k,k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}} + G_{k-1,k-1}^{\mathrm{T}}\Phi_{k,k-1}^{\mathrm{T}} + \Phi_{k,k-1}G_{k-1,k-1},$$
$$G_{k,k-1} = (\Phi_{k,k-1}G_{k-1,k-1} + L_{k-1,k-1})D_{k,k-1}^{\mathrm{T}},$$
$$L_{k,k-1} = D_{k,k-1}L_{k-1,k-1}D_{k,k-1}^{\mathrm{T}} + Q_{k-1},$$
$$P_{k,k} = [I - K_k^1H_k]\,P_{k,k-1},$$
$$G_{k,k} = [I - K_k^1H_k]\,G_{k,k-1},$$
$$L_{k,k} = L_{k,k-1} - K_k^2H_kG_{k,k-1},$$
$$\hat{x}_{0|0} = E\{x_0\},\ \hat{w}_0 = 0,$$
$$P_{0,0} = E\{[x_0 - E\{x_0\}][x_0 - E\{x_0\}]^{\mathrm{T}}\},$$
$$G_{0,0} = 0,\ L_{0,0} = 0.$$

3.3 闭环系统的卡尔曼滤波方法

卡尔曼滤波方法常常用于离散随机线性控制系统之中,这样的系统不仅包含模型噪声,而且含有控制输入项.其数学模型为

$$\Sigma_c:\ \begin{cases} x_k = \Phi_{k,k-1}x_{k-1} + \Gamma_{k,k-1}u_{k-1} + w_{k-1}, \\ y_k = H_kx_k + v_k. \end{cases}$$

其中 u_k 为 r 维控制输入向量;$\Gamma_{k,k-1}$ 为 $n \times r$ 阶矩阵,叫做**控制分布矩阵**.除此之外,其余各符号含义同前所述,并满足 1.1.3 小节的假设.这时对系统 Σ_c 来说,它的卡尔曼滤波器方程为

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k[y_k - H_k\hat{x}_{k|k-1}],$$
$$\hat{x}_{k|k-1} = \Phi_{k,k-1}\hat{x}_{k-1|k-1} + \Gamma_{k,k-1}u_{k-1},$$
$$K_k = P_{k,k-1}H_k^{\mathrm{T}}[H_kP_{k,k-1}H_k^{\mathrm{T}} + R_k]^{-1},$$
$$P_{k,k-1} = \Phi_{k,k-1}P_{k-1,k-1}\Phi_{k,k-1}^{\mathrm{T}} + Q_{k-1},$$
$$P_{k,k} = [I - K_kH_k]\,P_{k,k-1},$$

$$\hat{x}_{0|0} = E\{x_0\},$$
$$P_{0,0} = E\{[x_0 - E\{x_0\}][x_0 - E\{x_0\}]^{\mathrm{T}}\}.$$

4　连续时间随机线性系统的卡尔曼滤波

4.1　数学模型

4.1.1　状态方程和量测方程

连续时间随机线性系统的状态方程和量测方程一般可写为

$$\Sigma:\quad \dot{x}(t) = A(t)x(t) + w(t),$$
$$y(t) = C(t)x(t) + v(t),$$

其中$x(t)$表示 n 维随机状态向量；$y(t)$表示 m 维随机量测向量；$w(t)$表示 n 维系统模型噪声；$v(t)$表示 m 维系统量测噪声；$A(t)$为 $n\times n$ 阶矩阵，它的每个元素都是 t 的分段连续函数；$C(t)$表示 $m\times n$ 阶矩阵，它的每个元素都是 t 的分段连续函数.

令 $n\times n$ 阶矩阵 $\boldsymbol{\Phi}(t,t_0)$满足微分方程

$$\dot{\boldsymbol{\Phi}}(t,t_0) = A(t)\boldsymbol{\Phi}(t,t_0),$$
$$\boldsymbol{\Phi}(t_0,t_0) = I,$$

则称 $\boldsymbol{\Phi}(t,t_0)$为系统 Σ 的状态转移矩阵，它具有在 1.1.2 小节所叙述的性质.

4.1.2　基本假设

(1)假设系统 Σ 的初始状态 $x(t_0)$是一个正态分布的随机向量，其均值和协方差有穷，即

$$E\{x(t_0)\} = \bar{x}_0,\quad E\{x(t_0)x^{\mathrm{T}}(t_0)\} = P_0.$$

(2)假设系统 Σ 的模型噪声 $w(t)$是一个均值为零、协方差有穷的独立正态随机过程，或称零均值的白噪声过程，即对于任意 $t\geqslant\tau\geqslant 0$，有

$$E\{w(t)\} = 0,$$
$$E\{w(t)w^{\mathrm{T}}(\tau)\} = Q(t)\delta(t-\tau),$$

其中 $Q(t)$是 $n\times n$ 阶矩阵，且对称非负定，它的每个元都是 t 的分段连续函数；$\delta(t)$是脉冲函数，具有如下性质：

$$\int_{-\infty}^{\infty}\delta(t)\mathrm{d}t = 1,$$

且

$$\delta(t) = \begin{cases}\infty & (t=0);\\ 0 & (t\neq 0).\end{cases}$$

(3)假设系统 Σ 的量测噪声 $v(t)$也是一个均值为零、协方差有穷的独立正态

随机过程,即对于任意 $t \geqslant \tau \geqslant 0$,有

$$E\{\boldsymbol{v}(t)\} = 0,$$
$$E\{\boldsymbol{v}(t)\boldsymbol{v}^{\mathrm{T}}(\tau)\} = \boldsymbol{R}(t)\delta(t-\tau),$$

其中$\boldsymbol{R}(t)$是 $n \times n$ 阶对称正定矩阵,它的每个元都是 t 的分段连续函数.

(4)假设系统 Σ 的初始状态向量、模型噪声与量测噪声两两相互独立,即对于任意 $t,\tau \geqslant 0$,有

$$E\{\boldsymbol{x}(t)\boldsymbol{w}^{\mathrm{T}}(\tau)\} = 0,$$
$$E\{\boldsymbol{x}(t)\boldsymbol{v}^{\mathrm{T}}(\tau)\} = 0,$$
$$E\{\boldsymbol{w}(t)\boldsymbol{v}^{\mathrm{T}}(\tau)\} = 0.$$

通常称满足上述性质的系统 Σ 为正态系统.

4.1.3 问题的叙述

对于连续时间随机线性系统来说,最优状态估计问题的提法是:给定量测数据 $\{\boldsymbol{y}(t); t_0 \leqslant t \leqslant \tau\}$,寻找系统 Σ 的状态向量 $\boldsymbol{x}(t)$的估计 $\widehat{\boldsymbol{x}}(t|\tau)$,这个估计应具有下列三个性质:

1° $\widehat{\boldsymbol{x}}(t|\tau)$是无偏的,即 $E\{\widehat{\boldsymbol{x}}(t|\tau)\} = E\{\boldsymbol{x}(t)\}$;

2° $\widehat{\boldsymbol{x}}(t|\tau)$是线性的,即它是 $\boldsymbol{y}(t)$的线性函数;

3° $\widehat{\boldsymbol{x}}(t|\tau)$是最小方差估计,即它使得

$$E\{[\boldsymbol{x}(t) - \widehat{\boldsymbol{x}}(t)]^{\mathrm{T}}[\boldsymbol{x}(t) - \widehat{\boldsymbol{x}}(t)]\}$$

达到极小.

定义 1 如果满足上述三条性质的估计 $\widehat{\boldsymbol{x}}(t|\tau)$存在,则称它为线性无偏最小方差估计.

定义 2 若 $t > \tau$,则 $\widehat{\boldsymbol{x}}(t|\tau)$称为 $\boldsymbol{x}(t)$的预测估计,

定义 3 若 $t = \tau$,则 $\widehat{\boldsymbol{x}}(t|\tau)$称为 $\boldsymbol{x}(t)$的滤波估计,

定义 4 若 $t < \tau$,则 $\widehat{\boldsymbol{x}}(t|\tau)$称为 $\boldsymbol{x}(t)$的平滑估计.

卡尔曼滤波研究的是最优滤波估计问题.

4.2 卡尔曼滤波器

4.2.1 卡尔曼滤波器方程

连续时间线性随机系统 Σ 的卡尔曼滤波器可由下式给出:

$$\dot{\widehat{\boldsymbol{x}}}(t|t) = [\boldsymbol{A}(t) - \boldsymbol{K}(t)\boldsymbol{C}(t)]\widehat{\boldsymbol{x}}(t|t) + \boldsymbol{K}(t)\boldsymbol{y}(t), \tag{4-1}$$
$$\boldsymbol{K}(t) = \boldsymbol{P}(t)\boldsymbol{C}^{\mathrm{T}}(t)\boldsymbol{R}^{-1}(t), \tag{4-2}$$

$$\dot{P}(t)=A(t)P(t)+P(t)A^{T}(t)-P(t)C^{T}(t)R^{-1}(t)C(t)P(t)+Q(t),\tag{4-3}$$

$$\widehat{x}(0|0)=\overline{x}_0,\quad P(t_0)=P_0.$$

其中 $K(t)$ 为卡尔曼滤波器的最优增益矩阵.

4.2.2 卡尔曼滤波器的稳定性

定义 5 假设连续时间随机线性系统为 Σ,如果存在正常数 α_1,β_1 和 $\sigma>0$,使得对于一切 $t\geqslant 0$,都有

$$\alpha_1 I\leqslant\int_t^{t+\sigma}\Phi(t,\tau)Q(\tau)\Phi^{T}(t,\tau)d\tau\leqslant\beta_1 I,$$

则称该系统是一致完全能控的.

定义 6 假设连续时间随机线性系统为 Σ,如果存在正常数 α_2,β_3 和 $\sigma>0$,使得对于一切 $t\geqslant 0$,有

$$\alpha_2 I\leqslant\int_t^{t+\sigma}\Phi^{T}(\tau,t)C^{T}(\tau)R^{-1}(\tau)C(\tau)\Phi(\tau,t)d\tau\leqslant\beta_2 I,$$

则称该系统是一致完全能观测的.

定理 1 假设连续时间随机线性系统 Σ 是一致完全能控和一致完全能观测的,则系统 Σ 的卡尔曼滤波器具有如下性质:

1° 对于任意 t_0 时刻的初始协方差矩阵 $P_0\geqslant 0$,矩阵里卡蒂方程(4-3)式都有唯一对称非负定解 $P(t;t_0,P_0)$,而且当 $t\geqslant\sigma$ 时,$P(t;t_0,P_0)$ 总是正定的.

2° 对于任意 t_0 时刻的初始协方差矩阵 $P_0\geqslant 0$,当 $t\geqslant\sigma$ 时,$P(t;t_0,P_0)$ 是一致有界的,即存在正常数 α_3 和 β_3,使得对于一切 $t\geqslant 0$,有

$$\alpha_3 I\leqslant P(t;t_0,P_0)\leqslant\beta_3 I.$$

3° 对于任意 $P_0\geqslant 0$,有

$$\lim_{t_0\to-\infty}P(t;t_0,P_0)=\widetilde{P}(t)>0.$$

4° 最优滤波器方程(4-1)式是大范围一致渐近稳定的,即存在两个正常数 α 和 β,使得对于一切 $t>\tau\geqslant 0$,有

$$\|\Psi(t,\tau)\|\leqslant\alpha\exp(-\beta(t-\tau)),$$

其中 $\Psi(t,\tau)$ 为方程(4-1)式的状态转移矩阵,即 $\Psi(t,\tau)$ 满足下列微分方程:

$$\dot{\Psi}(t,\tau)=[A(t)-K(t)C(t)]\Psi(t,\tau),$$
$$\Psi(t_0,t_0)=I.$$

定理 2 假设系统 Σ 是定常的,即 $A(t)=A$, $C(t)=C$ 都是常值矩阵,模型噪声和量测噪声都是平稳的,即 $Q(t)=Q\geqslant 0$ 和 $R(t)=R>0$ 都是常值矩阵,并且存在矩阵 Γ,使得 $Q=\Gamma\Gamma^{T}$.如果$[A,\Gamma]$能控,$[A,C]$能观测,那么里卡蒂方程

$$AP+PA^{T}-PC^{T}R^{-1}CP+Q=0,\tag{4-4}$$

有唯一对称正定解 P,并且矩阵 $A-PC^{T}R^{-1}C$ 的所有特征值都有负实部.如果取

$$K=PC^{T}R^{-1},$$

则滤波器

$$\hat{x}(t|t)=(A-KC)\hat{x}(t|t)+Ky(t)$$

是渐近稳定的,这时它被称为系统 Σ 的稳态卡尔曼滤波器或常增益卡尔曼滤波器.

注 定理 2 中的假设[A,Γ]能控和[A,C]能观测分别可减弱为[A,Γ]能稳和[A,C]能检测,这时定理 2 仍然成立,只是代数里卡蒂方程(4-4)式只有唯一对称非负定解.

4.3 广义卡尔曼滤波

4.3.1 数学模型与基本假设

给定一个由下列微分方程描述的随机非线性系统:

$$\dot{x}(t)=f(x(t))+w(t), \tag{4-5}$$

$$y_k(t)=h(x_k)+v_k, \tag{4-6}$$

其中$x(t)$表示 n 维状态向量;y_k 表示 m 维量测向量;x_k 表示$x(t)$在 t_k 时刻的采样值;$w(t)$表示 n 维模型噪声向量;v_k 表示 m 维量测噪声向量;$f(\cdot)$表示状态向量$x(t)$的 n 维矢值函数;$h(\cdot)$表示向量 x_k 的 m 维矢值函数.

假设 $w(t)$是一个均值为零、协方差为

$$E\{w(t)w^{T}(\tau)\}=Q(t)\delta(t-\tau)$$

的正态白噪声过程;$\{v_k;k=1,2,\cdots\}$是一个均值为零、协方差为

$$E\{v_kv_j^{T}\}=R_k\delta_{kj}$$

的正态白噪声序列,这里 $Q(t)$和 R_k 分别为 $n\times n$ 阶对称非负定矩阵和 $m\times m$ 阶对称正定矩阵.另外假设 $E\{w(t)v_k^{T}\}=0$,对于一切 $t\geqslant 0$ 和 $k=1,2,\cdots$成立;$w(t)$,v_k都与初始状态x_0 相互独立,即

$$E\{w(t)x_0^{T}\}=0 \quad (\forall\, t\geqslant 0),$$

$$E\{v_kx_0^{T}\}=0 \qquad (\forall\, k=1,2,\cdots).$$

同时,还假设矢值函数 $f(x(t))$和 $h(x_k)$,对于每个状态变量的偏导数都存在.

现在的问题是在给出量测数据 $y_1,y_2,y_3,\cdots,y_k$ 以后,求状态向量 x_k 的估计 $\hat{x}_{k|k}$.

4.3.2 广义卡尔曼滤波器方程

利用线性化方程可得广义卡尔曼滤波器方程为

$$\hat{x}_{k|k}=\hat{x}_{k|k-1}+K_k[y_k-h(\hat{x}_{k|k-1})],$$

$$\hat{x}_{k|k-1}=\hat{x}_{k-1|k-1}+f(\hat{x}_{k-1|k-1})\Delta t,$$

$$K_k=P_{k,k-1}H_k^{T}[H_kP_{k,k-1}H_k^{T}+R_k]^{-1},$$

$$P_{k,k-1} = \Phi_{k,k-1} P_{k-1,k-1} \Phi_{k,k-1}^{\mathrm{T}} + Q_{k-1} \Delta t,$$
$$P_{k,k} = [I - K_k H_k] P_{k,k-1},$$
$$\hat{x}_{0|0} = E\{x_0\},$$
$$P_{0,0} = E\{[x_0 - E\{x_0\}][x_0 - E\{x_0\}]^{\mathrm{T}}\}.$$

其中 Δt 为采样时间间隔；$\Phi_{k,k-1} = I + F_{k,k-1}\Delta t$,

$$F_{k,k-1} = \left.\frac{\partial f(x_{k-1})}{\partial x_{k-1}}\right|_{x=\hat{x}_{k-1|k-1}},$$
$$H_k = \left.\frac{\partial h(x_k)}{\partial x_k}\right|_{x=\hat{x}_{k|k-1}}.$$

参 考 文 献

1 贾沛璋，朱征桃. 最优估计及其应用. 北京：科学出版社，1984.

2 中国科学院数学所概率组编著. 离散时间系统滤波的数学方法. 北京：国防工业出版社，1975.

3 王恩平，崔毅. 线性控制系统理论在惯性导航系统中的应用. 北京：科学出版社，1984.

·经济数学卷·

第 15 篇

系统辨识

编　者　萧德云
审校者　方崇智

目　录

引　　言

系统辨识是研究如何建立动态系统数学模型的一种理论和方法.在许多情况下,要建立一个系统的数学模型不是一件容易的事,尤其对那些具体的物理对象或工程系统,由于它们的内在机理比较复杂,而且系统的测量数据通常含有噪声,系统的建模就变得更为困难.系统辨识就是应此而形成的一门学科.它研究如何从含有噪声的测量数据中提取被研究对象的数学模型,或者说研究如何利用辨识技术从被研究的系统、对象或过程的复杂因果关系中,尽可能准确地确立它们之间的定量依存关系.

在控制工程中,系统辨识定义为"根据系统的输入输出数据,从给定的模型类中寻找与所研究的系统等价的模型".这一定义指出了系统的动态特性必然表现在它的变化着的输入输出数据之中.辨识就是从这变化的输入和输出数据中,利用数学方法提炼出系统的数学模型来的.一般说来,辨识建模是个复杂的实验测试和数据统计处理过程,通过辨识获得的模型也只是系统动态特性在某种准则意义下的一种近似.近似的程度取决于人们对系统先验知识的认识程度和对数据集合性质的了解,以及所采用的模型结构、辨识输入信号的性质、数据的预处理和准则函数的选择等因素.

系统辨识作为协同控制理论去解决实际控制工程问题的一种有力手段已被人们所公认.它是一门有明显实用价值的学科,涉及到的知识面比较宽,尤其对随机过程和线性代数的知识要求高.

1　基本概念

进行控制系统分析、设计时,需要知道控制对象的数学模型,在经典控制理论中,通常把正弦波信号或阶跃信号加到系统上,通过观测系统的输出来测定系统的动态特性模型.在现代控制理论中,系统数学模型的确定比较复杂,需要进行大量的数据处理和统计分析才能得到系统的动态特性模型.这些都是系统辨识研究的范畴.

粗略地说,系统辨识是根据测量数据来估计系统的数学模型的方法,而状态估计则是根据已知数学模型来估计系统的状态的方法,二者互为逆问题.

目前,控制理论的应用几乎不能没有系统辨识和状态估计的支持,系统辨识和状态估计又必须以控制理论为基础,三者构成了现代控制工程中三个互相渗透的学科领域.

1.1 系统建模

1.1.1 模型概述

控制理论和其他数理统计的最优化方法要应用于实际问题时，首先就需要知道系统的数学模型.所谓“模型”就是把关于实际系统的本质的部分信息简缩成有用的描述形式.它用来描述系统的运动规律，是系统的一种客观写照，是分析、预报、控制系统行为特性的基础.

对实际系统来说，建模必须要有明确的目的.一般说来，只能按照建模的目的去建立一种能近似描述系统的模型，而想建立包含系统各种因素在内的模型是不可能的.如果模型的输出和实际系统的输出“几乎处处”相等，那么应该说这种模型就是满意的了.要求模型越准确，模型就会越复杂.相反，适当降低模型的准确度要求，只考虑系统的主要因素，模型就可以简单点.这就是说，建立实际系统模型时，存在着准确性和复杂性这对矛盾，找出二者的折衷办法是建模的关键.

1.1.2 建模方法

常用的建模方法有两种：

1. 机理建模

机理建模也称理论建模.这种方法通过对系统的分析，搞清系统的运动规律，运用一些已知的定律、定理和原理，如物料平衡方程、能量平衡方程、传热传质原理和化学动力学原理等，来建立能准确反映系统内在机理的模型.

2. 辨识建模

辨识建模也称实验建模.这种建模方法的依据是系统的动态特性必然表现在输入输出数据之间的关系上，因此，可利用输入输出数据所提供的信息，建立与系统外特性等价的模型.

就某种意义上说，辨识建模较机理建模有一定的优越性，它无需深入了解系统的内在变化规律，然而，辨识建模需要设计一个合理的实验，以便获得系统所含的最大信息量，这点往往是困难的.因此，实际应用时，两种方法可以互相补充、取长补短.机理已知的部分可以用理论建模的方法，未知的部分则采用辨识建模的方法.通常把机理建模称作“白箱”问题，把辨识建模称作“黑箱”问题，二者的结合称作“灰箱”问题.

1.2 系统辨识

系统辨识是一种统计实验的建模方法，它只关心系统的外特性，把系统看做“黑箱”，根据“黑箱”表现出来的输入和输出信息，建立与“黑箱”特性等价的模型.或者说系统辨识是用一个模型来表示客观系统本质特征的一种演算，并把对客观系统的理解用模型的形式表示出来的方法.它并不期望获得一个与实际系统完全

吻合的数学描述，要的只是一个可适合于应用的模型．可见，辨识包含有三个基本要素——数据、模型类和准则．数据是辨识的基础，模型类是辨识的选择范围，准则是辨识的优化目标．

辨识就是按照一个准则在一组模型类中选择一个与数据拟合得最好的模型．总而言之，辨识的实质就是从一组模型类中选择一个模型，按照某种准则，使之能最好地拟合所关心的实际系统的动态特性．

1.3　辨识问题的描述

在控制工程领域中，多数场合都采用扎德（L. A. Zadeh）对系统辨识的定义："辨识就是在输入和输出数据的基础上，从一组给定的模型类中，确定一个与所测系统等价的模型"．根据这个定义，下面的几项工作在系统辨识中是非常重要的：

（1）输入信号的设计；

（2）输入和输出数据的测定；

（3）用于衡量实际系统和模型等价性的准则函数的确定；

（4）模型结构的确定；

（5）参数估计算法的选择；

（6）模型验证．

当然，这些工作是互相关联的，不能分开来考虑．

输入信号的设计必须满足一定的条件，以保证系统是可辨识的．也就是说，所设计的输入信号必须能充分激励系统的所有模态，使有关系统的动态特性信息尽可能正确地反映在输出数据上．

输入和输出数据的测定要考虑采样时间的选择、数据的预处理、坏数据的剔除和数据的零均值化等问题，尽可能保证数据集合的统计性质是平稳的或是拟平稳的．

系统和模型的等价性是通过引入具有物理意义的评价函数或称准则函数来衡量的．在给定的模型类中，当模型 $\mathscr{M}_0$ 使准则函数最小时，就称模型 $\mathscr{M}_0$ 与实际系统是等价的．这意味着辨识问题可以表达成寻求使准则函数最小的模型 $\mathscr{M}_0$ 的最优化问题．

模型结构的确定取决于模型的使用目的及有关的先验知识．

参数估计是系统辨识的一个阶段，当给定的模型用参数模型描述时，辨识问题便归结为模型参数优化问题，参数估计算法就是求这种优化问题的解的方法．

模型验证的目的是检查辨识模型的合适度，一般可以通过对所得的模型或基于模型所设计的控制系统进行模拟评价．

1.4　辨识的基本原理

辨识就是根据系统的测量数据，在某种准则意义下，对未知系统的模型结构和参数进行估计的方法．准则是用来判定模型是否接近实际系统的标准，可用模型与

系统偏差的泛函来表示.这个泛函称作准则函数,记作

$$J = \sum_{k=1}^{L} l(\varepsilon(k)),$$

其中 $l(\cdot)$ 是定义在 $(0, L)$ 区间上模型与系统误差 $\varepsilon(k)$ 的函数,通常称它为损失函数.多数情况下损失函数都取误差的平方函数,即

$$l(\varepsilon(k)) = \varepsilon^2(k).$$

误差 $\varepsilon(k)$ 应该广义地理解为模型与系统的一种"偏差",或称残差(residual).它可以是模型输出误差,也可以是模型输入误差或模型广义误差.输出误差的比较点位于模型的输出端,这时辨识问题可归结为非线性优化问题,辨识算法较复杂.输入误差的比较点在模型输入端,相应要求模型必须是可逆的.广义误差的比较点既不在输出端也不在输入端,通常要选择合适的比较点,使准则函数关于模型参数空间是线性的,这时模型参数优化问题比较简单,许多辨识算法都采用这种误差准则.

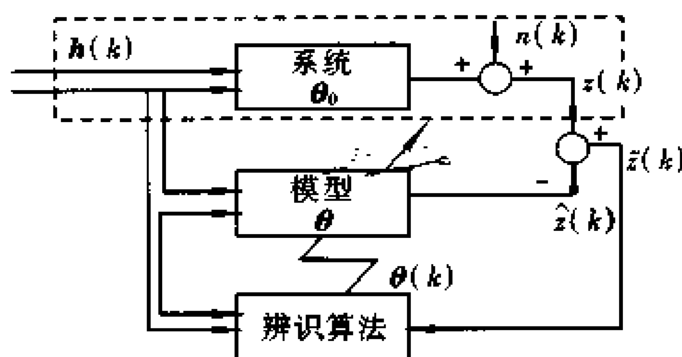


图 1-1

辨识的基本原理如图 1-1 表示,它是一个动态的调整动作过程,以实际系统的输出为参考,不断调整模型的结构和参数,使模型与实际系统输出之间的偏差尽可能小.

图 1-1 所示的虚线框内为待辨识的动态系统,其输出变量可以表示成另外一些变量在时间或空间离散点上取值的线性组合,或者说系统的模型可以表示成最小二乘格式.在 k 时刻,根据前一时刻的模型参数估计值 $\hat{\boldsymbol{\theta}}(k-1)$,可以计算模型的输出,也就是系统的输出预报值

$$\hat{z}(k) = \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1),$$

其中 $\boldsymbol{h}(k)$ 为模型输入数据向量,其元素是可测且相互间线性独立的.同时又可计算系统的输出预报误差,或称新息(innovation),即

$$\tilde{z}(k) = z(k) - \hat{z}(k),$$

其中 $z(k)$ 为系统的输出

$$\tilde{z}(k) = \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}_0 + n(k),$$

式中 $\boldsymbol{\theta}_0$ 为模型参数真值,$n(k)$ 为零均值噪声.然后将新息 $\tilde{z}(k)$ 反馈给辨识算法,在某种准则条件下,计算出 k 时刻的模型参数估计值 $\hat{\boldsymbol{\theta}}(k)$,并据此更新模型参数.这样不断迭代下去,直至对应的准则函数达到最小值为止.这时模型的输出也在该准则意义下最好地逼近系统的输出值,再通过比较不同模型结构下的准则函数,便可获得所需的模型.假设进行到 L 时刻获得了最终辨识模型,记作 $\hat{\boldsymbol{\theta}}(L)$,则

可进一步求得残差，即最终辨识模型与实际系统输出之间存在的偏差

$$\varepsilon(k) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(L) \qquad (k = 1,2,\cdots,L)$$

和相应的准则函数

$$J = \sum_{k=1}^{L} \varepsilon^2(k).$$

注意，残差与新息是两种不同的概念．前者用来衡量最终辨识模型与实际系统输出之间存在的偏差；后者表示辨识过程中模型与实际系统输出之间存在的动态偏差．

1.5　辨识的步骤及其分类

系统辨识从实验设计到获得最终模型，一般要经历如下步骤：

(1) 根据辨识的目的，利用先验知识，初步确定模型结构；

(2) 设计辨识实验方案，获取系统输入输出数据；

(3) 确定准则函数，完成模型参数和结构辨识；

(4) 通过模型检验，获得最终模型．

这些辨识步骤是密切关联而不是孤立的，其关系如图 1-2 所示．

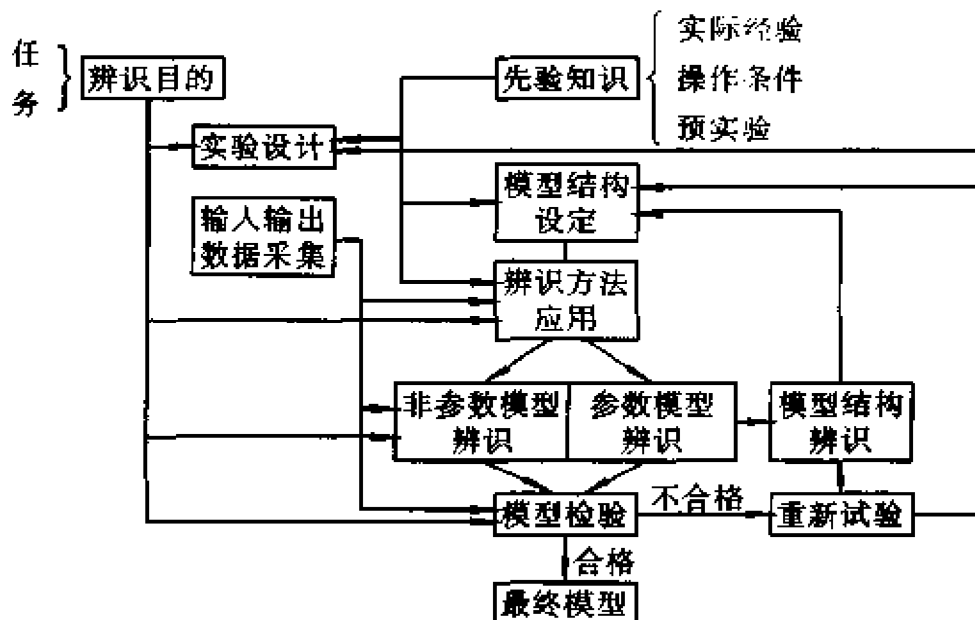


图 1-2

目前系统辨识技术已得到了迅速的发展，形成了各种各样的辨识方法．按误差准则分类有输出误差法、输入误差法和广义误差法；按算法形式分类有批处理算法和递推算法；按使用方式分类有离线辨识方法和在线辨识方法；按模型性质分类有非参数模型辨识方法(或称经典的辨识方法)和参数模型辨识方法(或称现代的辨识方法)．根据不同的辨识原理，参数模型辨识方法可归纳成三类：

(1) 最小二乘类参数辨识方法，其基本思想是通过极小化如下准则函数来估计模型参数，即

$$J(\hat{\theta}) = \sum_{k=1}^{L} \varepsilon^2(k) \Big|_{\hat{\theta}} \to \min,$$

其中 $\varepsilon(k)$ 代表模型输出与系统输出的偏差. 典型的方法有最小二乘法、增广最小二乘法、辅助变量法、广义最小二乘法等.

(2) 梯度校正参数辨识方法，其基本思想是沿着准则函数负梯度方向逐步修正模型参数，使准则函数达到最小，如随机逼近法.

(3) 概率密度逼近参数辨识方法，其基本思想是使输出 z 的条件概率密度 $p(z \mid \theta)$ 最大限度地逼近条件 θ_0 下的概率密度 $p(z \mid \theta_0)$，即

$$p(z \mid \hat{\theta}) \xrightarrow{\max} p(z \mid \theta_0).$$

典型的方法是极大似然法.

本篇重点讨论这三类辨识方法，对近年来发展起来的一些新的辨识方法，如模糊辨识方法[2,10]、神经网络辨识方法、小波辨识方法、遗传辨识方法等，由于篇幅所限，不作论述.

2　系统模型描述

在系统工程和控制工程领域中，研究诸如工业装置、生产过程、社会经济等实际系统的设计、控制、预报和评价时，需要一个与其目的相适应的数学模型，用以描述系统的动态行为. 在多数情况下，这种系统模型都可简化为线性时不变集中参数模型.

2.1　系统描述

系统描述是辨识的出发点，由此便可大致确定辨识问题的类型、辨识的应用范围和辨识所用的方法等. 一个系统的描述根据人们的需要可区分为时域描述和频域描述两种方式.

2.1.1　时域描述

1. 脉冲响应



图 2-1

考虑一个输入变量为 $u(t)$、输出变量为 $y(t)$ 的系统，如图 2-1 所示.

定义 1　如果系统响应只依赖于输入信号，不依赖于绝对时间，则系统称作时不变系统.

定义 2　如果系统对于输入信号线性组合的响应等于各个输入信号响应的线性组合，则系统称为线性系统.

定义 3　如果系统每一时刻的输出响应只依赖于到此时刻为止的输入，则系统称作因果系统．

一个线性时不变因果系统可以用脉冲响应来描述：

$$y(t)=\int_0^{\infty}g(\tau)u(t-\tau)\mathrm{d}\tau .$$

当脉冲响应$\{g(\tau),\tau\in(0,\infty)\}$和输入信号$u(s),s\leqslant t$已知时，系统完全可由脉冲响应表征．如果系统采样时间T足够小，输入信号在采样时间内可视为保持常值$u(t)=u(k),\forall kT\leqslant t<(k+1)T$，则系统可以用脉冲响应的离散形式来近似描述：

$$y(k)=\sum_{l=0}^{\infty}g(l)u(k-l), \tag{2-1}$$

其中

$$g(l)=\int_{lT}^{(l+1)T}g(\tau)\mathrm{d}\tau .$$

$\{g(l),l=0,1,\cdots\}$称为系统的脉冲响应序列．只要输入信号在采样时间内变化不大，脉冲响应序列就可以很好地描述系统．

2. 噪声

在实际应用中，系统总会受到不可观测的各种各样的噪声干扰．这些噪声通常都是随机的，记作$\{e(k)\}$，称作噪声序列．描述它们的基本数字特征定义为：

(1) 均值　$\mu_e(k)=E\{e(k)\}$；

(2) 均方值　$\psi_e^2(k)=E\{e^2(k)\}$；

(3) 方差　$\sigma_e^2(k)=E\{(e(k)-\mu_e(k))^2\}$；

(4) 相关函数　$R_e(k,k+l)=E\{e(k)e(k+l)\}$；

(5) 协方差　$C_e(k,k+l)=E\{(e(k)-\mu_e(k))(e(k+l)-\mu_e(k+l))\}$．

以上这些基本数字特征存在如下关系：

$$\begin{cases}\psi_e^2(k)=R_e(k,k),\\ \sigma_e^2(k)=R_e(k,k)-\mu_e^2(k),\\ C_e(k,k+l)=R_e(k,k+l)-\mu_e(k)\mu_e(k+l).\end{cases}$$

下面是一些有关随机序列的重要定义．

定义 4　如果一个随机序列的统计性质不随时间变化而变化，则称它为平稳随机序列．

定义 5　如果随机序列的一阶矩和二阶矩不随时间变化而变化，即$E\{e(k)\}=\mu_e$及$E\{e(k)e(k+l)\}=R_e(l)$，则称它为宽平稳随机序列．

定义 6　如果

$$\begin{cases}E\{e(k)\}=\mu_e(k) & (|\mu_e(k)|\leqslant c_1,\forall k),\\ E\{e(k)e(k+l)\}=R_e(k,k+l) & (|R_e(k,k+l)|\leqslant c_2,\forall k,l),\\ \overline{E}\{e(k)e(k+l)\}=\lim\limits_{L\to\infty}\dfrac{1}{L}\sum\limits_{k=1}^{L}E\{e(k)e(k+l)\}=R_e(l) & (\forall l),\end{cases}$$

则称它为拟平稳随机序列．

定义 7　如果$C_{ew}(l)=\overline{E}\{[e(k)-\mu_e(k)][w(k+l)-\mu_w(k+l)]\}=0$，则

$\{e(k)\}$ 和 $\{w(k)\}$ 是两个不相关随机序列.

定义8 拟平稳序列矩的计算定义为

$$\overline{E}\{f(e(k))\} = \lim_{L\to\infty}\frac{1}{L}\sum_{k=1}^{L}E\{f(e(k))\}.$$

如果不考虑概率框架,仅用一个实现来计算统计矩,则有

$$\overline{E}\{f(e(k))\} = \lim_{L\to\infty}\frac{1}{L}\sum_{k=1}^{L}f(e(k)),$$

这种利用一个实现去计算矩的概念属于各态遍历性范畴.

3. 噪声描述

定义9 如果一个随机噪声序列 $\{v(k)\}$ 是两两不相关的,对应的自相关函数为

$$R_v(l) = \sigma_v^2\delta_l \qquad (l = 0, \pm 1, \pm 2, \cdots),$$

其中 δ_l 为克罗内克(Z. Kronecker)符号,

$$\delta_l = \begin{cases}1 & (l = 0),\\ 0 & (l \neq 0),\end{cases}$$

则这种随机噪声序列称为**白噪声序列**.

白噪声序列的谱密度为常数,等于 σ_v^2,或者说等于白噪声的方差.

定理1 设随机平稳噪声序列 $\{e(k)\}$ 的谱密度 $S_e(\omega)$ 是 ω 的实函数,或是 $\cos\omega$ 的有理函数,那么必定存在一个渐近稳定的线性环节,使得如果环节的输入是白噪声,那么环节的输出是噪声序列 $\{e(k)\}$.

根据定义9和定理1,当系统受到噪声干扰时,则系统可描述成如图2-2所示的系统.噪声 $e(k)$ 是输入、输出及系统各方面干扰的综合,它可描述成

$$e(k) = \sum_{l=0}^{\infty}h(l)v(k-l), \tag{2-2}$$

且噪声 $e(k)$ 具有如下统计性质:

1° 均值 $E\{e(k)\} = \sum_{l=0}^{\infty}h(l)E\{v(k-l)\} = 0,$

2° 相关函数 $\begin{cases}E\{e(k)e(k+l)\} = \sigma_v^2\sum_{r=0}^{\infty}h(r)w(r+l),\\ h(r) = 0 \qquad (\forall r < 0).\end{cases}$

其中 $h(l), l \in (0, \infty)$,为噪声模型的脉冲响应序列.

4. 系统时域描述



图 2-2

如图2-2所示的系统可描述成:

$$z(k) = \sum_{l=0}^{\infty}g(l)u(k-l) + e(k). \tag{2-3}$$

引入迟延算子Z,记 $Z^{-1}u(k) = u(k-1)$,则(2-1)式可写成

$$y(k) = \sum_{l=0}^{\infty}g(l)u(k-l) = \left(\sum_{l=0}^{\infty}g(l)Z^{-l}\right)u(k) = G(Z^{-1})u(k),$$

其中 $G(Z^{-1}) = \sum_{l=0}^{\infty} g(l)Z^{-l}$,称作系统传递函数,也就是系统模型.

类似地,可把(2-2)式写成

$$e(k) = \sum_{l=0}^{\infty} h(l)v(k-l) = \left(\sum_{l=0}^{\infty} h(l)Z^{-l}\right)v(k) = H(Z^{-1})v(k),$$

其中 $H(Z^{-1}) = \sum_{l=0}^{\infty} h(l)Z^{-l}$,称作噪声传递函数,也就是噪声模型.

这样,线性时不变系统(2-3)式的基本描述可以写成

$$z(k) = G(Z^{-1})u(k) + H(Z^{-1})v(k), \tag{2-4}$$

称之为**线性时不变模型**,其中 $u(k)$ 和 $z(k)$ 分别是系统输入、输出变量;$v(k)$ 是均值为零、方差为 σ_v^2 的不相关随机噪声.

2.1.2 频域描述

1. 频率响应

定义 10 由(2-1)式描述的系统,其频率响应定义为

$$G(j\omega) = \frac{Y(j\omega)}{U(j\omega)},$$

其中 $U(j\omega)$,$Y(j\omega)$ 分别为系统输入和输出变量的傅里叶(J.B.Fourier)变换.频率响应 $G(j\omega)$ 与脉冲响应 $\{g(l)\}$ 的关系如下:

$$G(j\omega) = \sum_{l=0}^{\infty} g(l)\exp(-j\omega l).$$

定义 11 由(2-2)式描述的噪声频率响应 $H(j\omega)$ 与噪声脉冲响应 $\{h(l)\}$ 的关系定义为

$$H(j\omega) = \sum_{l=0}^{\infty} h(l)\exp(-j\omega l).$$

2. 谱密度

假定一个无限的信号序列为 $\{x(k), k = 1,2,\cdots\}$,则其自谱密度函数 $S_x(\omega)$ 与自相关函数 $R_x(l)$ 之间存在下述关系:

$$S_x(\omega) = \sum_{l=-\infty}^{+\infty} R_x(l)\exp(-j\omega l),$$

而 $\{x(k), k = 1,2,\cdots\}$ 与另一个信号序列 $\{y(k), k = 1,2,\cdots\}$ 的互谱密度函数 $S_{xy}(j\omega)$ 与互相关函数 $R_{xy}(l)$ 之间存在下述关系:

$$S_{xy}(j\omega) = \sum_{l=-\infty}^{+\infty} R_{xy}(l)\exp(-j\omega l).$$

谱密度描述了随机序列的平均功率分布.

3. 周期图

定义 12 假定一个有限的信号序列为 $\{x(k), k = 1,2,\cdots,L\}$,则其傅里叶变换定义为

$$X_L(\mathrm{j}\omega_i)=\sum_{k=1}^{L}x(k)\exp(-\mathrm{j}\omega_i k)\quad\left(\omega_i=\frac{2\pi i}{L},\ i=1,2,\cdots,L\right),$$

那么信号$\{x(k)\}$的周期图(periodgram)可写成

$$\begin{aligned}I_x(\omega_i)&=S_{x,L}(\omega_i)=\sum_{l=1}^{L}R_x(l)\exp(-\mathrm{j}\omega_i l)\\&=\frac{1}{L}\|X_L(\mathrm{j}\omega_i)\|^2\quad\left(\omega_i=\frac{2\pi i}{L},\ i=1,2,\cdots,L\right).\end{aligned}\tag{2-5}$$

根据帕塞瓦尔(Parseval)定理有

$$\sum_{i=1}^{L}I_x(\omega_i)=\sum_{k=1}^{L}x^2(k),$$

这说明信号$\{x(k)\}$周期图的代数和等于信号能量的代数和，或者说信号的能量可以分解为不同频率信号的能量贡献的代数和．

定理 2　设$\{x(k),k=1,2,\cdots\}$是一个拟平稳随机序列，则谱密度$S_x(\omega)$与周期图$I_x(\omega_i)$存在如下关系：

$$\lim_{L\to\infty}E\{I_x(\omega_i)\}=S_x(\omega).$$

其中

$$\begin{cases}I_x(\omega_i)=\dfrac{1}{L}\|X_L(\mathrm{j}\omega_i)\|^2 & \left(\omega_i=\dfrac{2\pi i}{L},i=1,2,\cdots,L\right),\\ X_L(\mathrm{j}\omega_i)=\displaystyle\sum_{k=1}^{L}x(k)\exp(-\mathrm{j}\omega_i k) & \left(\omega_i=\dfrac{2\pi i}{L},i=1,2,\cdots,L\right),\\ S_x(\omega)=\displaystyle\sum_{l=-\infty}^{+\infty}R_x(l)\exp(-\mathrm{j}\omega l), & \\ R_x(l)=\overline{E}\{x(k)x(k+l)\}. &\end{cases}$$

定理 2 给出了周期图与谱密度之间的关系．谱密度是在无限数据条件下获得的，但在实际应用中，总是利用周期图来计算谱密度．

4. 系统频域描述

利用谱密度函数，系统(2-4)式的频域描述可写成

$$\begin{cases}S_z(\omega)=\|G(\mathrm{j}\omega)\|^2S_u(\omega)+\|H(\mathrm{j}\omega)\|^2\sigma_v^2,\\ S_{zu}(\mathrm{j}\omega)=G(\mathrm{j}\omega)S_u(\omega).\end{cases}\tag{2-6}$$

如果(2-4)式是多变量系统，则其频域可描述为

$$\boldsymbol{S}_z(\omega)=\boldsymbol{G}(\mathrm{j}\omega)\boldsymbol{S}_u(\omega)\boldsymbol{G}^{\mathrm{T}}(\mathrm{j}\omega)+\boldsymbol{H}(\mathrm{j}\omega)\boldsymbol{\Sigma}_v(\omega)\boldsymbol{H}^{\mathrm{T}}(\mathrm{j}\omega),$$

其中$\boldsymbol{\Sigma}_v$为噪声协方差阵．

周期图和谱密度一样，也是用来描述随机序列的平均功率分布的，不同的是谱密度需要无限长的数据，而周期图用的是有限长的数据．周期图不像谱密度那样，在线性系统中的传递是线性的，无法导出(2-6)式那种结果，即

$$\begin{cases}I_z(\omega_i)\neq\|G(\mathrm{j}\omega_i)\|^2I_u(\omega_i)+\|H(\mathrm{j}\omega)\|^2\sigma_v^2,\\ I_{zu}(\mathrm{j}\omega_i)\neq G(\mathrm{j}\omega_i)I_u(\omega_i).\end{cases}$$

下面的定理间接描述了周期图在线性系统中的传递关系．

定理 3 假定一个严格稳定的线性系统为

$$y(k) = G(Z^{-1})u(k) \quad (\text{条件}: \sum_{l=0}^{\infty} l \mid g(l) \mid < \infty),$$

其中 $G(Z^{-1}) = \sum_{l=0}^{\infty} g(l) Z^{-l}$,且系统的输入是有界的,即

$$\mid u(k) \mid < c, \qquad \forall k, \text{则有}$$

$$Y_L(\mathrm{j}\omega_i) = G(\mathrm{j}\omega_i) U_L(\mathrm{j}\omega_i) + Q_L(\mathrm{j}\omega_i).$$

其中

$$\begin{cases} Y_L(\mathrm{j}\omega_i) = \sum_{k=1}^{L} y(k)\exp(-\mathrm{j}\omega_i k) & (\omega_i = \frac{2\pi i}{L}, i = 1,2,\cdots,L), \\ U_L(\mathrm{j}\omega_i) = \sum_{k=1}^{L} u(k)\exp(-\mathrm{j}\omega_i k) & (\omega_i = \frac{2\pi i}{L}, i = 1,2,\cdots,L), \\ \| Q_L(\mathrm{j}\omega_i) \| \leqslant 2c \sum_{l=0}^{\infty} l \mid g(l) \mid & (\omega_i = \frac{2\pi i}{L}, i = 1,2,\cdots,L). \end{cases}$$

由定理 3 和(2-5) 式,可求得周期图在线性系统中的传递关系 .

2.2 噪声模型

定理 4 设平稳随机序列$\{e(k)\}$均值为零,谱密度 $S_e(\omega)$ 是 $\cos\omega$ 的有理函数,则必定存在两个稳定的首 1 迟延算子多项式:

$$\begin{cases} C(Z^{-1}) = 1 + c_1 Z^{-1} + c_2 Z^{-2} + \cdots + c_{n_c} Z^{-n_c}, \\ D(Z^{-1}) = 1 + d_1 Z^{-1} + d_2 Z^{-2} + \cdots + d_{n_d} Z^{-n_d}, \end{cases}$$

使得

$$S_e(\omega) = \frac{\sigma_v^2 D(\mathrm{j}\omega) D^*(\mathrm{j}\omega)}{2\pi C(\mathrm{j}\omega) C^*(\mathrm{j}\omega)},$$

其中 $C^*(\mathrm{j}\omega)$ 和 $D^*(\mathrm{j}\omega)$ 是对应的共轭多项式 .

定理 4 给出重要的谱分解概念,它为系统(2-4) 式的噪声描述提供了一种标准的表达形式:

$$e(k) = H(Z^{-1})v(k) = \frac{D(Z^{-1})}{C(Z^{-1})} v(k). \tag{2-7}$$

这为噪声提供了很好的建模框架 .(2-7) 式所描述的噪声模型通常可分成如下几种类型。

(1) 自回归(auto regression,简记 AR) 模型

$$C(Z^{-1}) e(k) = v(k);$$

(2) 滑动平均(moving average,简记 MA) 模型

$$e(k) = D(Z^{-1}) v(k);$$

(3) 自回归滑动平均(auto regression moving average,简记 ARMA) 模型

$$C(Z^{-1})e(k) = D(Z^{-1})v(k).$$

2.3 线性时不变系统模型

线性时不变模型是辨识最常用的一类模型．下面的各类线性时不变模型是(2-4)式中的 $G(Z^{-1})$ 和 $H(Z^{-1})$ 取不同形式时的具体表现．

2.3.1 方程误差模型

1．ARX 模型

模型： $A(Z^{-1})z(k) = B(Z^{-1})u(k) + v(k).$

参数向量： $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{\mathrm{T}}.$

模型阶次： $n_a, n_b.$

预报器： $\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - A(Z^{-1})z(k) + B(Z^{-1})u(k).$

注意，若 $n_a = 0$，则 ARX 模型退化为 FIR(finited impluse response) 模型．

2．ARMAX 模型

模型： $A(Z^{-1})z(k) = B(Z^{-1})u(k) + D(Z^{-1})v(k).$

参数向量： $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}.$

模型阶次： $n_a, n_b, n_d.$

预报器： $\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - \dfrac{A(Z^{-1})}{D(Z^{-1})}z(k) + \dfrac{B(Z^{-1})}{D(Z^{-1})}u(k).$

注意，若以 $\Delta z(k) = z(k) - z(k-1)$ 代替模型中的 $z(k)$，则 ARMAX 模型[1]称为 ARIMAX 模型．

3．DA(dynamic adjustement) 模型

模型： $A(Z^{-1})z(k) = B(Z^{-1})u(k) + \dfrac{1}{C(Z^{-1})}v(k).$

参数向量： $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c})^{\mathrm{T}}.$

模型阶次： $n_a, n_b, n_c.$

预报器： $\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - A(Z^{-1})C(Z^{-1})z(k) + B(Z^{-1})C(Z^{-1})u(k).$

4．ARARMAX 模型

模型： $A(Z^{-1})z(k) = B(Z^{-1})u(k) + \dfrac{D(Z^{-1})}{C(Z^{-1})}v(k).$

参数向量： $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}.$

模型阶次： $n_a, n_b, n_c, n_d.$

预报器： $\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - \dfrac{A(Z^{-1})C(Z^{-1})}{D(Z^{-1})}z(k) + \dfrac{B(Z^{-1})C(Z^{-1})}{D(Z^{-1})}u(k).$

2.3.2 输出误差模型

1．测量误差模型

模型：　$z(k) = \frac{B(Z^{-1})}{A(Z^{-1})}u(k) + v(k).$

参数向量：　$\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{\mathrm{T}}.$

模型阶次：　$n_a, n_b.$

预报器：　$\hat{z}(k \mid \boldsymbol{\theta}) = \frac{B(Z^{-1})}{A(Z^{-1})}u(k).$

2. 博克斯－詹金斯(Box-Jenkins)模型

模型：　$z(k) = \frac{B(Z^{-1})}{A(Z^{-1})}u(k) + \frac{D(Z^{-1})}{C(Z^{-1})}v(k).$

参数向量：　$\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}.$

模型阶次：　$n_a, n_b, n_c, n_d.$

预报器：　$\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - \frac{C(Z^{-1})}{D(Z^{-1})}z(k) + \frac{B(Z^{-1})C(Z^{-1})}{A(Z^{-1})D(Z^{-1})}u(k).$

2.3.3　一般结构模型

模型：　$A(Z^{-1})z(k) = \frac{B(Z^{-1})}{F(Z^{-1})}u(k) + \frac{D(Z^{-1})}{C(Z^{-1})}v(k).$

参数向量：　$\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c}, d_1, d_2, \cdots, d_{n_d}, f_1, f_2, \cdots, f_{n_f})^{\mathrm{T}}.$

模型阶次：　$n_a, n_b, n_c, n_d, n_f.$

预报器：　$\hat{z}(k \mid \boldsymbol{\theta}) = z(k) - \frac{A(Z^{-1})C(Z^{-1})}{D(Z^{-1})}z(k) + \frac{B(Z^{-1})C(Z^{-1})}{F(Z^{-1})D(Z^{-1})}u(k).$

注 1　以上各类模型中，其相应的迟延算子多项式取下述形式：

$$\begin{cases} A(Z^{-1}) = 1 + a_1 Z^{-1} + a_2 Z^{-2} + \cdots + a_{n_a} Z^{-n_a}, \\ B(Z^{-1}) = 1 + b_1 Z^{-1} + b_2 Z^{-2} + \cdots + b_{n_b} Z^{-n_b}, \\ C(Z^{-1}) = 1 + c_1 Z^{-1} + c_2 Z^{-2} + \cdots + c_{n_c} Z^{-n_c}, \\ D(Z^{-1}) = 1 + d_1 Z^{-1} + d_2 Z^{-2} + \cdots + d_{n_d} Z^{-n_d}, \\ F(Z^{-1}) = 1 + f_1 Z^{-1} + f_2 Z^{-2} + \cdots + f_{n_f} Z^{-n_f}. \end{cases}$$

注 2　以上各类模型的预报器可以定义成一个统一的框架结构，其中预报值关于参数是线性的，即$\hat{z}(k \mid \boldsymbol{\theta}) = \boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{\theta}$，在 k 时刻，$\boldsymbol{\varphi}(k)$ 是已知的，其元素可以是数据的某种变换．

在应用以上各类模型时，需注意如下几个问题：

(1) 以上各类模型是用于描述 SISO(single input/single output) 系统的，对于 MIMO(multiple input/multiple output) 系统，一般采用状态空间模型描述．

(2) 以上各类模型的预报器是辨识算法求新息所必须的．无论是 SISO 系统，还是 MIMO 系统，它们均可转化成如下的状态空间模型：

$$\begin{cases} \boldsymbol{h}(k+1,\boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{h}(k,\boldsymbol{\theta}) + \boldsymbol{G}(\boldsymbol{\theta})\bar{\boldsymbol{u}}(k), \\ \hat{\boldsymbol{z}}(k \mid \boldsymbol{\theta}) = \boldsymbol{H}(\theta)\boldsymbol{h}(k,\boldsymbol{\theta}). \end{cases} \tag{2-8}$$

上式描述的是 MIMO 系统输出预报器的一般表达式,SISO 系统是它的特例 .(2-8)式预报模型的输入 $\bar{\boldsymbol{u}}(k) \in \mathbf{R}^{(m+r)\times 1}$ 由系统的输入 $\boldsymbol{u}(k) \in \mathbf{R}^{r\times 1}$($r$ 为系统输入维数)和输出 $z(k) \in \mathbf{R}^{m\times 1}$($m$ 为系统输出维数)组成,即

$$\bar{\boldsymbol{u}}(k) = [z^{\mathrm{T}}(k)\ \boldsymbol{u}^{\mathrm{T}}(k)]^{\mathrm{T}};$$

$\hat{z}(k \mid \boldsymbol{\theta}) \in \mathbf{R}^{m\times 1}$ 为模型输出预报;$\boldsymbol{h}(k,\boldsymbol{\theta})$ 为模型状态变量;$\boldsymbol{F}(\boldsymbol{\theta})$,$\boldsymbol{G}(\boldsymbol{\theta})$ 和 $\boldsymbol{H}(\boldsymbol{\theta})$ 为模型参数 $\boldsymbol{\theta} = (\theta_1,\theta_2,\cdots,\theta_N)^{\mathrm{T}}$ 的矩阵函数,它们的构成取决于具体的模型结构 .

(3) 辨识算法的组成很大程度上依赖于模型输出预报值及其关于参数 $\boldsymbol{\theta}$ 的一阶梯度 . 下述模型是(2-8)式的进一步描述,它同时描述了预报值及预报值关于参数梯度的动态关系:

$$\begin{cases} \boldsymbol{x}(k+1,\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}(k,\boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{\theta})\bar{\boldsymbol{u}}(k), \\ \begin{bmatrix} \hat{z}(k \mid \boldsymbol{\theta}) \\ \mathrm{col}\boldsymbol{\Psi}^{\mathrm{T}}(k,\boldsymbol{\theta}) \end{bmatrix} = \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{x}(k,\boldsymbol{\theta}), \end{cases} \tag{2-9}$$

其中

$$\boldsymbol{x}(k+1,\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{h}(k,\boldsymbol{\theta}) \\ \mathrm{col}\boldsymbol{\Phi}^{\mathrm{T}}(k,\boldsymbol{\theta}) \end{bmatrix},$$

$$\boldsymbol{\Phi}(k,\boldsymbol{\theta}) = \frac{\partial \boldsymbol{h}(k,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

$$\boldsymbol{\Psi}(k,\boldsymbol{\theta}) = \frac{\partial \hat{z}(k,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

$$\boldsymbol{A}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{F}(\boldsymbol{\theta}) & \mathbf{0} & \cdots & \mathbf{0} \\ \dfrac{\partial \boldsymbol{F}(\boldsymbol{\theta})}{\partial \theta_1} & \boldsymbol{F}(\boldsymbol{\theta}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial \boldsymbol{F}(\boldsymbol{\theta})}{\partial \theta_N} & \mathbf{0} & \cdots & \boldsymbol{F}(\boldsymbol{\theta}) \end{bmatrix},$$

$$\boldsymbol{B}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{G}(\boldsymbol{\theta}) \\ \dfrac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \theta_N} \end{bmatrix},$$

$$\boldsymbol{C}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{H}(\boldsymbol{\theta}) & \mathbf{0} & \cdots & \mathbf{0} \\ \dfrac{\partial \boldsymbol{H}(\boldsymbol{\theta})}{\partial \theta_1} & \boldsymbol{H}(\boldsymbol{\theta}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial \boldsymbol{H}(\boldsymbol{\theta})}{\partial \theta_N} & \mathbf{0} & \cdots & \boldsymbol{H}(\boldsymbol{\theta}) \end{bmatrix}.$$

上述讨论的线性时不变模型是对动态系统常用的描述形式,本篇主要研究这

类模型的辨识问题．线性时变和非线性系统模型也是辨识常用的模型．如果(2-4)式模型的传递函数还与时间有关，则(2-4)式可写成

$$z(k) = G(k, Z^{-1})u(k) + H(k, Z^{-1})v(k),$$

该模型便成为线性时变模型了．线性时变模型仅仅引入了时变传递函数，其他细节及预报器的结构与线性时不变模型的基本一样，所用的辨识方法也类似．关于非线性系统的描述及其辨识方法，由于篇幅的限制，这里不再作进一步的论述．

3　最小二乘法

第2章讨论了各种类型的系统模型结构，这些模型都可用来描述系统的动态特征．不论是哪类系统，模型响应的确切性都是由模型参数决定的．第3章至第7章主要研究在适当的实验条件下，通过对系统响应的观测来估计模型参数的方法．这些方法统称模型参数辨识方法．

目前，已有很多种模型参数辨识方法，各种方法都有两种算法形式：① **批处理算法**，利用一批观测数据，一次计算或经反复迭代，以获得模型参数的估计值．② **递推算法**，在上次模型参数估计值 $\hat{\boldsymbol{\theta}}(k-1)$ 的基础上，根据当前获得的数据进行修正，进而获得本次模型参数估计值 $\hat{\boldsymbol{\theta}}(k)$．广泛采用的递推算法形式为

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)\boldsymbol{h}(k-d)\tilde{z}(k). \tag{3-1}$$

其中 $\hat{\boldsymbol{\theta}}(k)$ 表示 k 时刻的模型参数估计值；$\boldsymbol{K}(k)$ 为算法的增益；$\boldsymbol{h}(k-d)$ 是由观测数据组成的输入数据向量，d 为整数；$\tilde{z}(k)$ 表示新息．

(3-1)式看起来很简单，然而随着所采用准则函数的不同，$\hat{\boldsymbol{\theta}}(k)$，$\boldsymbol{K}(k)$，$\boldsymbol{h}(k-d)$ 和 $\tilde{z}(k)$ 的确切含义和内容就会不同，由此形成的算法形式也是多种多样的．

在各种模型参数辨识方法中，最小二乘法是最基本的一种方法，其应用相当广泛．本章着重讨论最小二乘参数辨识方法及其各种变形．

3.1　最小二乘原理

定义1　设一个随机序列 $\{z(k), k \in (1,2,\cdots,L)\}$ 的均值是参数 $\boldsymbol{\theta}$ 的线性函数，即

$$E\{z(k)\} = \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta},$$

其中 $\boldsymbol{h}(k)$ 是可测的数据向量，那么利用随机序列的一个实现，使准则函数

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{L}[z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}]^2$$

达到极小的参数估计值 $\hat{\boldsymbol{\theta}}$ 称为 $\boldsymbol{\theta}$ 的**最小二乘估计**．

上述最小二乘原理表明，所谓未知参数估计问题，就是求参数估计值 $\hat{\boldsymbol{\theta}}$，使序

列的估计值尽可能地接近实际序列的问题，二者的接近程度用实际序列与序列估计值之差的平方和来度量．

假定系统的输入输出关系可以描述成如下的最小二乘格式：

$$z(k) = \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta} + n(k), \tag{3-2}$$

其中 $z(k)$ 为模型输出变量，$\boldsymbol{h}(k)$ 为输入数据向量，$\boldsymbol{\theta}$ 为模型参数向量，$n(k)$ 为零均值随机噪声．

为了求模型(3-2)式的参数估计值，可以利用上述最小二乘原理，根据已知的观测数据序列$\{z(k)\}$和$\{\boldsymbol{h}(k)\}$，极小化准则函数

$$J(\theta) = \sum_{k=1}^{L}[z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}]^2,$$

即可求得模型参数的最小二乘估计值 $\hat{\boldsymbol{\theta}}$，若其值使不同时刻观测值与估计值之间的误差的平方和达到最小值，则所得到的模型输出就认为能最好地逼近实际系统的输出．

3.2　最小二乘问题的解

考虑如下模型：

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + n(k). \tag{3-3}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$n(k)$ 是均值为零、方差为 σ_n^2 的随机噪声；$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式，可写成

$$\begin{cases} A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_{n_a}Z^{-n_a}, \\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_{n_b}Z^{-n_b}, \end{cases} \tag{3-4}$$

其中 n_a 和 n_b 为模型阶次．定义：

$$\begin{cases} \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}, \\ \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{\mathrm{T}} \end{cases}$$

将模型(3-3)式写成最小二乘格式

$$z(k) = \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta} + n(k). \tag{3-5}$$

对于 $k = 1,2,\cdots,L$（L 为数据长度），由(3-5)式可以构成如下线性方程组：

$$\boldsymbol{z}_L = \boldsymbol{H}_L\boldsymbol{\theta} + \boldsymbol{n}_L, \tag{3-6}$$

其中

$$\begin{cases} \boldsymbol{z}_L = (z(1), z(2), \cdots, z(L))^{\mathrm{T}}, \\ \boldsymbol{H}_L = \begin{bmatrix} \boldsymbol{h}^{\mathrm{T}}(1) \\ \boldsymbol{h}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}^{\mathrm{T}}(L) \end{bmatrix} = \begin{bmatrix} -z(1-1) & \cdots & -z(1-n_a) & u(1-1) & \cdots & u(1-n_b) \\ -z(2-1) & \cdots & -z(2-n_a) & u(2-1) & \cdots & u(2-n_b) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -z(L-1) & \cdots & -z(L-n_a) & u(L-1) & \cdots & u(L-n_b) \end{bmatrix}, \\ \boldsymbol{n}_L = (n(1), n(2), \cdots, n(L))^{\mathrm{T}}. \end{cases} \tag{3-7}$$

准则函数取

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{L} \Lambda(k)[z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}]^2, \tag{3-8}$$

其中 $\Lambda(k)$ 为加权因子，对所有的 k，$\Lambda(k)$ 都必须大于零．根据(3-7)式，准则函数又可写成

$$J(\boldsymbol{\theta}) = (\boldsymbol{z}_L - \boldsymbol{H}_L\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Lambda}_L(\boldsymbol{z}_L - \boldsymbol{H}_L\boldsymbol{\theta}), \tag{3-9}$$

其中 $\boldsymbol{\Lambda}_L$ 为加权矩阵，它是正定的对角阵，由加权因子 $\Lambda(k)$ 构成，

$$\boldsymbol{\Lambda}_L = \begin{bmatrix} \Lambda(1) & 0 & \cdots & 0 \\ 0 & \Lambda(2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Lambda(L) \end{bmatrix}.$$

显然，准则函数(3-9)式可用以衡量模型输出与实际系统输出的接近情况．极小化这个准则函数，即可求得模型的参数估计值，使模型的输出能最好地预报系统的输出．

当 $\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L$ 是正则矩阵时，模型(3-3)式的加权最小二乘解为

$$\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{z}_L. \tag{3-10}$$

通过极小化(3-9)式准则函数 $J(\boldsymbol{\theta})$，来求得模型参数估计值 $\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}$ 的方法称为加权最小二乘法，记作 **WLS**(weighted least squares algorithm)，对应的 $\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}$ 称为加权最小二乘估计值．如果加权矩阵取单位阵，即 $\boldsymbol{\Lambda}_L = \boldsymbol{I}$，则加权最小二乘解退化成普通最小二乘解：

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{z}_L. \tag{3-11}$$

这时的 $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ 称之为最小二乘估计值，对应的估计方法称作最小二乘法，记作 **LS**(least squares algorithm)．最小二乘法是加权最小二乘法的一种特例．

3.3 最小二乘估计的可辨识性

由(3-10)式和(3-11)式可知，最小二乘估计的可辨识条件为矩阵 $\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L$ 必须是非奇异的．这一要求与数据集是“提供信息”的，或辨识输入信号是“持续激励”的概念密切相关．

定义 2 称一个拟平稳数据集 D^{∞} 关于模型集 $\mathscr{M}$ 是“信息充足”的，如果在这个数据集合中，对于任意两个模型 $M_1(Z^{-1})$ 和 $M_2(Z^{-1})$，有

$$\overline{E}\{[(M_1(Z^{-1}) - M_2(Z^{-1}))\boldsymbol{D}(k)]^2\} = 0,$$

其中 $\boldsymbol{D}(k) = [u(k), z(k)]^{\mathrm{T}}$，也就是

$$M_1(\exp(-\mathrm{j}\omega)) = M_2(\exp(-\mathrm{j}\omega)) \qquad (\forall\,\omega).$$

定义 3 称一个拟平稳数据集 D^{∞} 是“提供信息”的，如果这个数据集关于由所有线性时不变模型组成的模型集 $\mathscr{M}$ 是“信息充足”的．

定理 1 如果 $D(k)=[u(k),z(k)]^{\mathrm{T}}$ 的谱矩阵

$$S_D(\mathrm{j}\omega)=\begin{bmatrix} S_u(\omega) & S_{uz}(\mathrm{j}\omega) \\ S_{zu}(\mathrm{j}\omega) & S_z(\omega) \end{bmatrix},$$

对于所有的 ω 是严格正定的,那么这个拟平稳数据集 D^{∞} 是"**提供信息**"的.

这种"信息充足"或"提供信息"的数据集概念与下面"持续激励"的输入信号概念是紧密联系的.

定义 4 设输入信号 $u(k)$ 是个拟平稳的随机信号,如果它的谱密度函数

$$S_u(\omega)>0 \qquad (\forall\omega),$$

则称 $u(k)$ 为"**持续激励**"信号.

定义 5 一个具有谱密度为 $S_u(\omega)$ 的拟平稳信号 $u(k)$ 称为 n 阶"持续激励"信号,若对于一切形如 $F_n(Z^{-1})=f_1Z^{-1}+f_2Z^{-2}+\cdots+f_nZ^{-n}$ 的滤波器,有如下关系式

$$\| F_n(\exp(-\mathrm{j}\omega))\|^2 S_u(\omega)\equiv 0,$$

也就是 $F_n(\exp(-\mathrm{j}\omega))\equiv 0$.

定理 2 设输入信号 $u(k)$ 是个拟平稳的随机信号,如果相关函数矩阵

$$R_u^n=\begin{bmatrix} R_u(0) & R_u(1) & \cdots & R_u(n-1) \\ \vdots & \vdots & \vdots & \vdots \\ R_u(n-1) & R_u(n-2) & \cdots & R_u(0) \end{bmatrix}$$

是非奇异的,则输入信号 $u(k)$ 是 n 阶"持续激励"信号.

假定有如下模型:

$$z(k)=G(Z^{-1})u(k)+H(Z^{-1})v(k), \tag{3-12}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量;$v(k)$ 为零均值白噪声;$G(Z^{-1})$ 和 $H(Z^{-1})$ 分别为过程和噪声模型,则当

$$\Delta H(\exp(-\mathrm{j}\omega))\equiv 0 \text{ 及 } \|\Delta G(\exp(-\mathrm{j}\omega))\|^2 S_u(\omega)\equiv 0$$

时,其中

$$\Delta H(\exp(-\mathrm{j}\omega))=H_1(\exp(-\mathrm{j}\omega))-H_2(\exp(-\mathrm{j}\omega)),$$

$$\Delta G(\exp(-\mathrm{j}\omega))=G_1(\exp(-\mathrm{j}\omega))-G_2(\exp(-\mathrm{j}\omega)),$$

模型(3-12)式生成的数据集必定是"信息充足"的.根据定义(3-4)式和 $\|\Delta G(\exp(-\mathrm{j}\omega))\|^2 S_u(\omega)\equiv 0$ 可知,模型(3-12)式生成的数据集是"信息充足"的条件是要求输入信号 $u(k)$ 必须是"持续激励"信号.

定义 6 如果辨识实验产生的数据集 D^{∞} 是"信息充足"的,则该辨识实验是"信息充足"的.

定理 3 假定模型(3-12)式中,过程模型 $G(Z^{-1})$ 为有理函数,即

$$G(Z^{-1})=\frac{b_1Z^{-1}+b_2Z^{-2}+\cdots+b_{n_b}Z^{-n_b}}{1+a_1Z^{-1}+a_2Z^{-2}+\cdots+a_{n_a}Z^{-n_a}},$$

则输入 $u(k)$ 为 n_a+n_b 阶"持续激励"信号的辨识实验是"信息充足"的.

推论 1 如果辨识输入信号是"持续激励"的,则开环辨识实验是"提供信息"的.

由以上定义和定理可知,为保证矩阵 $H_L^{\mathrm{T}}\Lambda_L H_L$ 是非奇异的,或者说使最小二乘估计是开环可辨识的,其充分必要条件是输入信号 $u(k)$ 必须是 $2n(n=\max(n_a,n_b))$

阶"持续激励"信号．该条件也就是最小二乘估计的开环可辨识性条件，可写成

$$R_u^{2n} > 0$$

或

$$\overline{U}_L^{\mathrm{T}} \overline{U}_L > 0.$$

其中相关函数矩阵 R_u^{2n} 按定理 2 构成，矩阵 $\overline{U}_L$ 按下式组成：

$$\begin{cases} \overline{U}_L = [\,F u_L \quad F^2 u_L \quad \cdots \quad F^{2n} u_L\,], \\ u_L = [\,u(1) \quad u(2) \quad \cdots \quad u(L)\,], \\ F = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{L\times L}, \\ n = \max(n_a, n_b). \end{cases}$$

3.4 最小二乘估计的几何解释

定义 7 设 y 是具有前二阶矩的 N 维随机向量，X 是适当维的随机矩阵．如果存在一个与 y 同维的随机向量 $\hat{y}$，它具备以下三个性质：

1° $\hat{y}$ 可以用 X 线性表示，即 $\hat{y} = a + Xb$，

2° $\hat{y}$ 是 y 的无偏估计，即 $E\{\hat{y}\} = E\{y\}$，

3° $(y - \hat{y})$ 与 X 正交，即 $E\{(y - \hat{y})^{\mathrm{T}} X\} = 0^{\mathrm{T}}$，

则称 $\hat{y}$ 是 y 在 X 空间上的正交投影．

根据定义 7，对模型(3-3)式的最小二乘估计和加权最小二乘估计可作如下几何解释：

如果模型(3-3)式的噪声 $n(k)$ 是零均值不相关随机噪声或称白噪声，又记

$$\begin{cases} H_L = [\,h_1 \quad h_2 \quad \cdots \quad h_N\,] \qquad (N = n_a + n_b), \\ \hat{z}_L = H_L \hat{\theta}_{\mathrm{LS}} \text{ 或 } \hat{z}_L = H_L \hat{\theta}_{\mathrm{WLS}}, \\ \varepsilon_L = z_L - \hat{z}_L, \end{cases}$$

则模型输出估计向量 $\hat{z}_L$ 就是系统输出向量 z_L 由线性不相关的数据向量 $h_1, h_2, \cdots, h_N$ 所张成的空间上的正交投影，或者说输出残差向量 ε_L 垂直于该空间．这时，对数据加权不会影响正交关系．

3.5 最小二乘估计的统计性质

(3-11)式和(3-10)式给出了模型(3-3)式的最小二乘估计和加权最小二乘估计，因式中数据矩阵 H_L 和输出向量 z_L 均具有随机性，故参数估计值 $\hat{\theta}_{\mathrm{LS}}$ 或 $\hat{\theta}_{\mathrm{WLS}}$ 亦

为随机向量．当模型的噪声满足一定条件时，它们具有如下一些优良的统计性质．

3.5.1 无偏性

无偏性是用来衡量参数估计值是否围绕真值波动的一个性质．

定理 4 如果模型(3-6)式噪声向量 $\boldsymbol{n}_L$ 是零均值且与数据矩阵 $\boldsymbol{H}_L$ 统计独立，则最小二乘参数估计值 $\hat{\boldsymbol{\theta}}_{LS}$ 或加权最小二乘参数估计值 $\hat{\boldsymbol{\theta}}_{WLS}$ 是无偏估计量，即

$$E\{\hat{\boldsymbol{\theta}}_{LS}\} = \boldsymbol{\theta}_0 \text{ 或 } E\{\hat{\boldsymbol{\theta}}_{WLS}\} = \boldsymbol{\theta}_0,$$

其中 $\boldsymbol{\theta}_0$ 为模型参数真值．

3.5.2 参数估计偏差的协方差阵性质

参数估计偏差的协方差阵是用来评价参数估计精度的一个重要依据．

定理 5 如果模型(3-6)式的噪声向量 $\boldsymbol{n}_L$ 是零均值且与数据矩阵 $\boldsymbol{H}_L$ 统计独立的随机向量，则加权最小二乘参数估计偏差 $\tilde{\boldsymbol{\theta}}_{WLS} = \boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{WLS}$ 的协方差阵可写成：

$$\operatorname{cov}\{\tilde{\boldsymbol{\theta}}_{WLS}\} = E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{\Sigma}_n\boldsymbol{\Lambda}_L\boldsymbol{H}_L(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L)^{-1}\},$$

其中 $\boldsymbol{\Sigma}_n$ 是噪声协方差阵，$\boldsymbol{\theta}_0$ 为模型参数真值．

推论 2 如果模型(3-6)式的噪声向量 $\boldsymbol{n}_L$ 是零均值白噪声，且加权矩阵取 $\boldsymbol{\Lambda}_L = \boldsymbol{I}$，则最小二乘参数估计偏差 $\tilde{\boldsymbol{\theta}}_{LS} = \boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{LS}$ 的协方差阵为

$$\operatorname{cov}\{\tilde{\boldsymbol{\theta}}_{LS}\} = \sigma_n^2 E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\},$$

其中 σ_n^2 是噪声方差，$\boldsymbol{\theta}_0$ 为模型参数真值．

推论 3 在定理 5 条件下，如果加权矩阵取 $\boldsymbol{\Lambda}_L = \boldsymbol{\Sigma}_n^{-1}$，则模型(3-6)式的参数估计值为

$$\hat{\boldsymbol{\theta}}_{MV} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{z}_L, \tag{3-13}$$

相应的参数估计偏差 $\tilde{\boldsymbol{\theta}}_{MV} = \boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{MV}$ 的协方差阵为

$$\operatorname{cov}\{\tilde{\boldsymbol{\theta}}_{MV}\} = E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}_L)^{-1}\}.$$

如果模型噪声 $n(k)$ 又同时服从正态分布，则(3-13)式给出的参数估计值偏差的方差将达到最小值．这时的参数估计称的最小方差估计，也称为**马尔可夫(A.A. Markov)估计**．显然，马尔可夫估计是加权最小二乘估计的一种特例．

3.5.3 一致性

一致性描述参数估计值的收敛情况．

定理 6 在推论 2 条件下，最小二乘参数估计是一致收敛的，即有

$$\lim_{L\to\infty}\hat{\boldsymbol{\theta}}_{LS} \longrightarrow \boldsymbol{\theta}_0 \quad (\text{W.P.1}).$$

3.5.4 有效性

有效性表明参数估计值偏差的协方差阵将达到下界．

定理 7 在推论 2 条件下,设模型噪声服从正态分布,则最小二乘参数估计值 $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ 是有效估计值,即参数估计值偏差的协方差阵达到克拉默-拉奥(Cramer-Rao)不等式的下界:

$$\mathrm{cov}\{\widetilde{\boldsymbol{\theta}}_{\mathrm{LS}}\} = \sigma_n^2 E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\} = \boldsymbol{M}^{-1},$$

其中 $\boldsymbol{M}$ 为费希尔(R.A.Fisher)**信息矩阵**

$$\boldsymbol{M} = E\left\{\left[\frac{\partial \ln p(z_L \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{\mathrm{T}}\left[\frac{\partial \ln p(z_L \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\Bigg|_{\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}}\right\}.$$

推论 4 在推论 3 条件下,设模型噪声服从正态分布,则马尔可夫参数估计 $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$ 也是有效估计,即

$$\mathrm{cov}\{\widetilde{\boldsymbol{\theta}}_{\mathrm{MV}}\} = E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}_L)^{-1}\} = \boldsymbol{M}^{-1}.$$

3.5.5 渐近分布性质

定理 8 在推论 2 条件下,设模型噪声服从正态分布,则最小二乘估计具有如下一些分布性质:

1° $\widetilde{\boldsymbol{\theta}}_{\mathrm{LS}} \sim N(\boldsymbol{\theta}_0, \sigma_n^2 E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\})$,

2° $\boldsymbol{\varepsilon}_L \sim N(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I})$,且 $E\{\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}\boldsymbol{\varepsilon}_L^{\mathrm{T}}\} = 0$,

3° $\dfrac{\boldsymbol{\varepsilon}_L^{\mathrm{T}}\boldsymbol{\varepsilon}_L}{\sigma_n^2} \xrightarrow{\text{a.s.}} \dfrac{\boldsymbol{n}_L^{\mathrm{T}}\boldsymbol{n}_L}{\sigma_n^2} \sim \chi^2(L - \dim\boldsymbol{\theta})$,

4° $\widetilde{\boldsymbol{\theta}}_i\left(\dfrac{p_i\boldsymbol{\varepsilon}_L^{\mathrm{T}}\boldsymbol{\varepsilon}_L}{L - \dim\boldsymbol{\theta}}\right)^{-\frac{1}{2}} \sim t(L - \dim\boldsymbol{\theta}) \quad (i = 1,2,\cdots,N)$,

其中 p_i 是矩阵$[\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L]^{-1}$第 i 个主对角线元素.

推论 5 在推论 3 条件下,设模型噪声服从正态分布,则马尔可夫参数估计值 $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$ 也服从正态分布,即

$$\widetilde{\boldsymbol{\theta}}_{\mathrm{MV}} \sim N(\boldsymbol{\theta}_0, \sigma_n^2 E\{(\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}{}_n\boldsymbol{H}_L)^{-1}\}).$$

3.5.6 噪声方差估计的性质

定理 9 在推论 2 条件下,模型噪声方差 σ_n^2 的估计值

$$\widehat{\sigma_n^2} = \frac{\boldsymbol{\varepsilon}_L^{\mathrm{T}}\boldsymbol{\varepsilon}_L}{L - \dim\boldsymbol{\theta}}$$

是无偏一致估计,其中 $\boldsymbol{\varepsilon}_L = z_L - \boldsymbol{H}_L\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$,定义为输出残差向量.

综上分析,最小二乘参数估计具有许多优良的统计性质,它们可以总结成一条重要的结论:对模型(3-3)式来说,如果模型噪声 $n(k)$ 是均值为零,且服从正态分布的白噪声,则模型参数 $\boldsymbol{\theta}$ 的最小二乘估计值是无偏、一致、有效估计.

3.6 最小二乘递推算法

所谓递推算法就是根据新的观测数据实时修正参数估计值,随着时间的推移,逐步获得满意的辨识结果的方法.

3.6.1 递推算法形式

在 $2n$ 阶"持续激励"输入信号的作用下,模型(3-3)式的加权最小二乘解为

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} &= [\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{H}_L]^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Lambda}_L\boldsymbol{z}_L \\ &= \Big(\sum_{i=1}^{L}\Lambda(i)\boldsymbol{h}(i)\boldsymbol{h}^{\mathrm{T}}(i)\Big)^{-1}\Big(\sum_{i=1}^{L}\Lambda(i)\boldsymbol{h}(i)z(i)\Big).\end{aligned}$$

其中 $\Lambda(i)$ 为加权因子.

记 k 时刻的参数估计值为

$$\hat{\boldsymbol{\theta}}(k) = \Big(\sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)\boldsymbol{h}^{\mathrm{T}}(i)\Big)^{-1}\Big(\sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)z(i)\Big), \tag{3-14}$$

令 $\overline{\boldsymbol{R}}(k) = \sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)\boldsymbol{h}^{\mathrm{T}}(i)$,利用

$$\overline{\boldsymbol{R}}(k-1)\hat{\boldsymbol{\theta}}(k-1) = \sum_{i=1}^{k-1}\Lambda(i)\boldsymbol{h}(i)z(i),$$

则有

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \overline{\boldsymbol{R}}^{-1}(k)\boldsymbol{h}(k)\Lambda(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \overline{\boldsymbol{R}}(k) = \overline{\boldsymbol{R}}(k-1) + \Lambda(k)\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k).\end{cases}$$

又设 $\boldsymbol{R}(k) = \frac{1}{k}\overline{\boldsymbol{R}}(k)$,可导出如下的加权最小二乘估计递推算法,记作 WRLS(weighted recursive least squares algorithm):

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \frac{1}{k}\boldsymbol{R}^{-1}(k)\boldsymbol{h}(k)\Lambda(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{R}(k) = \boldsymbol{R}(k-1) + \frac{1}{k}[\Lambda(k)\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k) - \boldsymbol{R}(k-1)].\end{cases} \tag{3-15}$$

置 $\boldsymbol{P}(k) = \frac{1}{k}\boldsymbol{R}^{-1}(k) = \Big(\sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)\boldsymbol{h}^{\mathrm{T}}(i)\Big)^{-1}$,并利用矩阵反演公式:

$$(\boldsymbol{A} + \boldsymbol{CBC}^{\mathrm{T}})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{C})\boldsymbol{C}^{\mathrm{T}}\boldsymbol{A}^{-1},$$

那么算法(3-15)式就可演变成下面所示的另一种递推算法形式:

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}(k)\Big(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) + \frac{1}{\Lambda(k)}\Big)^{-1}, \\ \boldsymbol{P}(k) = [\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k)]\boldsymbol{P}(k-1).\end{cases} \tag{3-16}$$

上面给出的两种加权最小二乘估计递推算法形式都是很常用的,其中(3-15)式多用于理论分析,而(3-16)式比较适用于在线计算.如果取加权因子 $\Lambda(k)=1$, $\forall K$,则两种加权最小二乘递推算法就变成普通的最小二乘递推算法,记作 **RLS**(recursive least squares algorithm).

由算法(3-16)式还可以推导出如下一些关于矩阵 $\boldsymbol{P}(k)$ 的递推关系:

$$\boldsymbol{P}(k)\boldsymbol{h}(k)=\boldsymbol{P}(k-1)\boldsymbol{h}(k)[1+\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)\Lambda(k)]^{-1},$$
$$\boldsymbol{P}(k-1)\boldsymbol{h}(k)=\boldsymbol{P}(k)\boldsymbol{h}(k)[1-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)\Lambda(k)]^{-1},$$
$$\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)=\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)[1+\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)\Lambda(k)]^{-1},$$
$$\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)=\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)[1-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)\Lambda(k)]^{-1},$$
$$\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}^{2}(k)\boldsymbol{h}(k)=\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}^{2}(k-1)\boldsymbol{h}(k)[1+\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)\Lambda(k)]^{-2},$$
$$\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}^{2}(k-1)\boldsymbol{h}(k)=\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}^{2}(k)\boldsymbol{h}(k)[1-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)\Lambda(k)]^{-2}.$$

另外,由于算法(3-16)式中的 $\boldsymbol{P}(k)$ 是一个对称、非增的矩阵.为了保证计算过程中 $\boldsymbol{P}(k)$ 矩阵始终是对称的,算法(3-16)式的第3式可采用下面的计算式,以保证不破坏 $\boldsymbol{P}(k)$ 矩阵的对称性:

$$\boldsymbol{P}(k)=\boldsymbol{P}(k-1)-\boldsymbol{K}(k)\boldsymbol{K}^{\mathrm{T}}(k)\Big[\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+\frac{1}{\Lambda(k)}\Big].$$

在计算过程中,增益矩阵 $\boldsymbol{K}(k)$ 可能产生误差,经过算法的迭代,造成误差传递和累积,最后将影响辨识算法的准确度.为此,可采用下面的计算式,以截断误差的传递,保证辨识精度:

$$\boldsymbol{P}(k)=[\boldsymbol{I}-\boldsymbol{K}(k)\boldsymbol{K}^{\mathrm{T}}(k)]\boldsymbol{P}(k-1)[\boldsymbol{I}-\boldsymbol{K}(k)\boldsymbol{K}^{\mathrm{T}}(k)]^{\mathrm{T}}+\boldsymbol{K}(k)\Lambda(k)\boldsymbol{K}^{\mathrm{T}}(k).$$

3.6.2 初始值的选择

递推算法(3-16)式的初始值一般可取为

$$\begin{cases}\boldsymbol{P}(0)=\alpha^{2}\boldsymbol{I},\\ \hat{\boldsymbol{\theta}}(0)=\boldsymbol{\varepsilon},\end{cases}$$

其中 α 为充分大实数,$\boldsymbol{\varepsilon}$ 为充分小实向量.这是因为(3-14)式可以写成

$$\hat{\boldsymbol{\theta}}(k)=\Big[\boldsymbol{P}^{-1}(0)+\sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)\boldsymbol{h}^{\mathrm{T}}(i)\Big]^{-1}\Big[\boldsymbol{P}^{-1}(0)\hat{\boldsymbol{\theta}}(0)+\sum_{i=1}^{k}\Lambda(i)\boldsymbol{h}(i)z(i)\Big].$$

显然,上式要与(3-14)式保持一致,在选择初始值时,必须使 $\boldsymbol{P}^{-1}(0)$ 和 $\hat{\boldsymbol{\theta}}(0)$ 都很小,应接近于0.

3.6.3 准则函数的递推计算

根据(3-16)式算法,可以导出残差 $\varepsilon(k)$ 与新息 $\tilde{z}(k)$ 的关系为

$$\varepsilon(k)=\frac{\tilde{z}(k)}{1+\Lambda(k)\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)}$$

或

$$\varepsilon(k)=[1-\Lambda(k)\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)]\tilde{z}(k).$$

由此可推出准则函数 $J(k)$ 的递推计算式为

$$J(k)=J(k-1)+\frac{\tilde{z}^2(k)}{h^{\mathrm{T}}(k)P(k-1)h(k)+\Lambda^{-1}(k)},$$

其中 $\tilde{z}(k)=z(k)-h^{\mathrm{T}}(k)\hat{\theta}(k-1)$,是 k 时刻的新息,它与 $k-1$ 时刻的参数估计值有关.

3.6.4　$P(k)$ 矩阵的基本性质

如果模型(3-3)式的输入是 $2n$ 阶"持续激励"信号,则矩阵 $P(k)$ 具有如下一些基本性质:

1° $P(k)$ 是对称、非奇异矩阵;

2° $\lambda_{\min}\{P^{-1}(k)\}\geqslant\lambda_{\min}\{P^{-1}(k-1)\}\geqslant\cdots\geqslant\lambda_{\min}\{P^{-1}(0)\}$,

其中 $\lambda_{\min}\{\cdot\}$ 表示矩阵的最小特征值;

3° $\lim\limits_{k\to\infty}\lambda_{\min}\{P^{-1}(k)\}=\infty$;

4° $\|P(k)\|=\dfrac{1}{|\lambda_{\min}\{P^{-1}(k)\}|}\xrightarrow[k\to\infty]{}0$;

5° $\lim\limits_{k\to\infty}P(k)=0$　(W.P.1).

3.7　最小二乘递推算法的收敛性

上面提到的两种最小二乘递推算法(3-15)式和(3-16)式都涉及时变差分方程的运算问题,其收敛性并不是无条件地成立的.

定理 10　考虑模型(3-3)式的最小二乘辨识问题,只有当噪声 $n(k)$ 为零均值白噪声时,递推算法(3-16)式给出的参数估计值 $\hat{\theta}(k)$ 才是一致收敛的,即有

$$\lim_{k\to\infty}\hat{\theta}(k)=\theta_0\qquad(\text{W.P.1.}),$$

其中 θ_0 为模型参数真值.

定理 10 的证明依靠以下一些事实:

(1) 由 $\tilde{\theta}(k)=P(k)P^{-1}(k-1)\tilde{\theta}(k-1)-K(k)n(k)$,可以得到

$$\tilde{\theta}(k)=\prod_{i=1}^{k}P(i)P^{-1}(i-1)\tilde{\theta}(0)-\sum_{j=1}^{k}\prod_{i=j+1}^{k}P(i)P^{-1}(i-1)P(j)h(j)n(j)$$

和

$$\tilde{\theta}(k)=\frac{1}{a^2}P(k)\tilde{\theta}(0)-P(k)\sum_{j=1}^{k}h(j)n(j);$$

(2) $\lim\limits_{k\to\infty}P(k)=0$　(W.P.1.);

(3) 由于噪声 $n(k)$ 是白噪声,故有

$$P(k)\sum_{j=1}^{k}h(j)n(j)=\frac{1}{k}\left[\frac{1}{k}\sum_{i=1}^{k}h(i)\Lambda(i)h^{\mathrm{T}}(i)\right]^{-1}k\left[\frac{1}{k}\sum_{i=1}^{k}h(i)n(i)\right]$$

$$\xrightarrow{k\to\infty}E\{h(i)\Lambda(i)h^{\mathrm{T}}(i)\}E\{h(i)n(i)\}=0.$$

定理 11 考虑模型(3-3)式的最小二乘辨识问题,如果噪声 $n(k)$ 为零均值白噪声,则递推算法(3-15)式给出的参数估计值 $\hat{\boldsymbol{\theta}}(k)$ 是一致收敛的,即有

$$\lim_{k\to\infty}\hat{\boldsymbol{\theta}}(k) = \boldsymbol{\theta}_0 \quad (\text{W.P.1.}),$$

其中 $\boldsymbol{\theta}_0$ 为模型参数真值.

定理 11 的证明思路如下:

(1) 算法(3-15)式的伴随微分方程为

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{\theta}_{\mathrm{D}}(\tau) = \boldsymbol{R}_{\mathrm{D}}^{-1}(\tau)\boldsymbol{f}(\boldsymbol{\theta}_{\mathrm{D}}),\\ \dfrac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{R}_{\mathrm{D}}(\tau) = \boldsymbol{G}(\boldsymbol{\theta}_{\mathrm{D}}) - \boldsymbol{R}_{\mathrm{D}}(\tau),\\ \boldsymbol{f}(\boldsymbol{\theta}_{\mathrm{D}}) = \lim\limits_{k\to\infty}E\{\boldsymbol{h}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}_{\mathrm{D}}(\tau))\},\\ \boldsymbol{G}(\boldsymbol{\theta}_{\mathrm{D}}) = \lim\limits_{k\to\infty}E\{\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k)\},\end{cases} \tag{3-17}$$

其中 $\tau = \sum_{i=1}^{k}\left(\dfrac{1}{i}\right)$.

(2) 考虑到 $n(k)$ 是零均值白噪声,则微分方程(3-17)式的平衡点 $\boldsymbol{\theta}_D^* = \boldsymbol{\theta}_0$.

(3) 构造微分方程(3-17)式的李雅普诺夫(Lyapunov)函数

$$V(\boldsymbol{\theta}_{\mathrm{D}}) = \frac{1}{2}E\{[\varepsilon(k,\boldsymbol{\theta}_{\mathrm{D}}) - n(k)]^2\},$$

其中

$$\varepsilon(k,\boldsymbol{\theta}_{\mathrm{D}}) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}_{\mathrm{D}}.$$

(4) 推导出

$$\frac{\mathrm{d}}{\mathrm{d}l}V(\boldsymbol{\theta}_{\mathrm{D}}) = -E\{[\boldsymbol{h}^{\mathrm{T}}(k)(\boldsymbol{\theta}_{\mathrm{D}} - \boldsymbol{\theta}_0)]^2\} \leqslant 0 \quad (\forall\,\boldsymbol{\theta}_{\mathrm{D}}).$$

(5) 由此可以得到微分方程(3-17)式的不变集为 $D_C = \{\boldsymbol{\theta}_0\}$,吸收域 D_A 为全平面.

(6) 微分方程(3-17)式的平衡点 $\boldsymbol{\theta}_D^* = \boldsymbol{\theta}_0$ 就是辨识算法(3-15)式参数估计值 $\hat{\boldsymbol{\theta}}(k)$ 的收敛点.

(7) 微分方程(3-17)式 $\boldsymbol{\theta}_{\mathrm{D}}(\tau)$ 的收敛轨迹就是辨识算法(3-15)式 $\hat{\boldsymbol{\theta}}(k)$ 的渐近收敛路径.

3.8 最小二乘递推算法的几种变形

最小二乘递推算法有多种不同的变形,常用的有七种:

(1) 基于数据所含的信息内容不同,对数据进行有选择性的加权;

(2) 在认为新近的数据更有价值的假设下,逐步丢弃过去的数据;

(3) 只用有限长度的数据;

(4) 加权方式既考虑平均特性又考虑跟踪能力;

(5) 在不同的时刻,重调协方差阵 $\boldsymbol{P}(k)$;

(6) 设法防止协方差阵 $\boldsymbol{P}(k)$ 趋于零;

(7) 有约束的最小二乘法.

其中第(2)、(4)、(5) 和(6) 种变形算法的重要性在于使最小二乘算法有可能用于时变系统,这就大大扩展了最小二乘法的应用范围.

3.8.1 选择性加权最小二乘算法

把加权最小二乘递推算法(3-16) 式改写成

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)),\\ \boldsymbol{K}(k) = \Lambda(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)(\Lambda(k)\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+1)^{-1},\\ \boldsymbol{P}(k) = [\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k)]\boldsymbol{P}(k-1).\end{cases} \tag{3-18}$$

算法中引进了加权因子,其目的是便于考虑观测数据的可信度.选择不同的加权方式对算法的性质会有影响,下面是几种特殊的选择.

(1) 当 $\Lambda(k)$ 取得很大时,在极限情况下,算法(3-18) 式就退化成正交投影算法.也就是说,当选择

$$\Lambda(k) = \begin{cases}\infty, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) \neq 0,\\ 0, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) = 0,\end{cases}$$

这样就构成了正交投影算法:

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)[z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)],\\ \boldsymbol{K}(k) = \dfrac{\boldsymbol{P}(k-1)\boldsymbol{h}(k)}{\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)},\\ \boldsymbol{P}(k) = [\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k)]\boldsymbol{P}(k-1).\end{cases} \tag{3-19}$$

算法的初始值取 $\boldsymbol{P}(0) = \boldsymbol{I}$ 及 $\hat{\boldsymbol{\theta}}(0) = \boldsymbol{\varepsilon}$(任给定值),当 $\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) = 0$ 时,令 $\boldsymbol{K}(k) = 0$.

定理 12 (3-19) 式正交投影辨识算法具有如下一些性质:

1° $\boldsymbol{P}(k)$ 是等幂矩阵,即 $\boldsymbol{P}^2(k) = \boldsymbol{P}(k)$;

2° $\boldsymbol{P}(0)\cdots\boldsymbol{P}(k) = \boldsymbol{P}(k)$;

3° $\boldsymbol{P}(k-1)\boldsymbol{h}(k)$ 是 $\boldsymbol{h}(1),\boldsymbol{h}(2),\cdots,\boldsymbol{h}(k)$ 的线性组合;

4° $\boldsymbol{P}(k)\boldsymbol{x} = \boldsymbol{0}$ 的充分必要条件是 $\boldsymbol{x}$ 为 $\boldsymbol{h}(1),\boldsymbol{h}(2),\cdots,\boldsymbol{h}(k)$ 的线性组合;

5° $\boldsymbol{P}(k-1)\boldsymbol{h}(k)$ 与 $\boldsymbol{h}(1),\boldsymbol{h}(2),\cdots,\boldsymbol{h}(k-1)$ 正交.

定理 12 表明,$\boldsymbol{P}(k-1)\boldsymbol{h}(k)$ 是 $\boldsymbol{h}(1),\boldsymbol{h}(2),\cdots,\boldsymbol{h}(k)$ 的线性组合,且与 $\boldsymbol{h}(1),\boldsymbol{h}(2),\cdots,\boldsymbol{h}(k-1)$ 正交.可见矩阵 $\boldsymbol{P}(k-1)$ 是实现这种正交投影的一种算子.

(2) 第(1) 种加权因子的选择显然是一种极端情况.由于要求无限的数值精度,且数据被简单分成接受或丢弃两种可能,因此算法的鲁棒性比较差.为了使算法具有较好的鲁棒性,可把第(1) 种加权因子的选择方案修改为:

$$\Lambda(k) = \begin{cases}\Lambda_1, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) \geqslant \epsilon,\\ \Lambda_2, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) < \epsilon,\end{cases}$$

其中 $\Lambda_1 \geqslant \Lambda_2 > 0$,$\epsilon$ 是指定的阈值.这时算法对数据作了不同的加权,但不排斥任

何数据 .

(3) 按下式选择加权因子,意味着它是过去数据信息量的一种度量:

$$\Lambda(k)=\begin{cases}\dfrac{\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)}{\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{h}(k)}, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{h}(k)\neq 0,\\ 0, & \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{h}(k)=0.\end{cases}$$

(4) 如果由噪声、建模不准确等因素引起的误差上界已知,则可按下式选择加权因子:

$$\Lambda(k)=\begin{cases}1, & \dfrac{[z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)]^2}{1+\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)}\geqslant\Delta^2>0,\\ 0, & \dfrac{[z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)]^2}{1+\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)}<\Delta^2>0.\end{cases}$$

3.8.2 遗忘因子算法

各种加权最小二乘算法只不过是对数据进行加权处理,过去的数据并没有被遗忘.随着时间的推移,新数据所提供的信息将被淹没在老数据的海洋之中.如果不衰减掉老数据所含的信息,新数据的信息就无法用上,算法就会慢慢失去修正能力.

遗忘因子算法通过对数据加遗忘因子的办法来降低老数据的信息量,为补充新数据的信息创造条件 . 取准则函数为

$$J(\boldsymbol{\theta})=\sum_{k=1}^{L}\mu^{L-k}[z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}]^2, \tag{3-20}$$

其中 μ 称为遗忘因子,取值为 $0<\mu<1$. 由极小化这个准则函数,可得到模型(3-3)式的参数辨识算法:

$$\hat{\boldsymbol{\theta}}_{\mathrm{FF}}=(\boldsymbol{H}_L^{*\mathrm{T}}\boldsymbol{H}_L^*)^{-1}\boldsymbol{H}_L^{*\mathrm{T}}\boldsymbol{z}_L^*,$$

其中

$$\begin{cases}\boldsymbol{z}_L^*=(\beta^{L-1}z(1),\beta^{L-2}z(2),\cdots,z(L))^{\mathrm{T}},\\ \boldsymbol{H}_L^*=\begin{bmatrix}\beta^{L-1}\boldsymbol{h}^{\mathrm{T}}(1)\\ \beta^{L-2}\boldsymbol{h}^{\mathrm{T}}(2)\\ \vdots\\ \boldsymbol{h}^{\mathrm{T}}(L)\end{bmatrix},\\ \beta^2=\mu.\end{cases}$$

这种参数辨识方法称作**遗忘因子法**,记作 FF(forgetting factor algorithm). 如果遗忘因子 $\mu=1$,则算法退化成普通最小二乘法 . 与推导加权最小二乘递推算法一样,同样可以推导出遗忘因子算法的递推计算形式:

$$\begin{cases}\hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\boldsymbol{K}(k)[z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)],\\ \boldsymbol{K}(k)=\boldsymbol{P}(k-1)\boldsymbol{h}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+\mu)^{-1},\\ \boldsymbol{P}(k)=\dfrac{1}{\mu}[\boldsymbol{I}-\boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k)]\boldsymbol{P}(k-1).\end{cases} \tag{3-21}$$

其中遗忘因子 μ 可按下面的原则取值：

(1) 若要求 T_c 步后数据衰减至 36%，则 $\mu = 1 - \frac{1}{T_c}$；

(2) μ 取作时变因子，即

$$\mu(k) = \mu_0\mu(k-1) + (1-\mu_0),$$

其中 $\mu_0 = 0.99, \mu(0) = 0.95$.

遗忘因子 μ 的取值大小对算法的性能会产生直接的影响．μ 值增加时，算法的跟踪能力下降，但算法的鲁棒性增强；μ 值减少时，算法的跟踪能力增强，但算法的鲁棒性减弱，对噪声更显得敏感．在实际应用中要折衷考虑这两方面的情况．

比较(3-8)式和(3-20)式两种准则函数后会发现，遗忘因子法似乎可以看成是加权最小二乘算法的一种特例，相当于加权因子 $\Lambda(k) = \mu^{L-k}$ 的情况．实际上这个观点是不对的，这两种算法不能简单地等同，主要的差别有：

(1) 加权方式不同　加权最小二乘法的各时刻权重是不相关联的，也不随时间变化而变化；遗忘因子法的各时刻权重是相关联的，满足

$$\Lambda(k) = \frac{1}{\mu}\Lambda(k-1)$$

关系，各时刻权重的大小随时间变化而变化，当前时刻的权重总为 1.

(2) 加权的效果不一样　加权最小二乘法获得的是系统的平均特性；遗忘因子法能实时跟踪系统的变化，具有跟踪能力．

(3) 算法的协方差矩阵 $\boldsymbol{P}(k)$ 的内容不一样　二者的关系为 $\boldsymbol{P}_{FF}(k) = \Lambda(k)\boldsymbol{P}_{WLS}(k)$.

和加权最小二乘递推算法一样，运用遗忘因子算法的准则函数(3-20)式也可实现递推计算．根据(3-21)式，可以导出遗忘因子算法下的残差 $\varepsilon(k)$ 与新息 $\tilde{z}(k)$ 的关系：

$$\varepsilon(k) = \frac{\mu\tilde{z}(k)}{\mu + \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)},$$

或

$$\varepsilon(k) = [1 - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}(k)]\tilde{z}(k).$$

由此可推出准则函数 $J(k)$ 的递推计算式：

$$J(k) = \mu\left(J(k-1) + \frac{\tilde{z}^2(k)}{\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) + \mu}\right),$$

其中 $\tilde{z}(k) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)$，是 k 时刻的新息，它与 $k-1$ 时刻的参数估计值有关．

3.8.3　限定记忆算法

加权最小二乘或遗忘因子递推算法在递推计算过程中数据长度是不断增长的，不论多老的数据都在起作用．也就是说，老数据所含的信息在算法中会不断累积，长期下去新数据的作用会被削弱．这种数据长度随 k 的增加不断增长的辨识

算法称做增长记忆法．另一类辨识算法叫做限定记忆法，它依赖于有限长度的数据，每增加一个新的数据信息，就要去掉一个老数据的信息，数据长度始终保持不变．这种方法的参数估计递推算法如下：

$$\begin{cases}\hat{\boldsymbol{\theta}}(k+1,k+L)=\hat{\boldsymbol{\theta}}(k,k+L)-\boldsymbol{K}(k+1,k+L)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k,k+L)),\\ \boldsymbol{K}(k+1,k+L)=\boldsymbol{P}(k,k+L)\boldsymbol{h}(k)(1-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k,k+L)\boldsymbol{h}(k))^{-1},\\ \boldsymbol{P}(k+1,k+L)=(\boldsymbol{I}+\boldsymbol{K}(k+1,k+L)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k,k+L),\\ \hat{\boldsymbol{\theta}}(k,k+L)=\hat{\boldsymbol{\theta}}(k,k+L-1)+\boldsymbol{K}(k,k+L)(z(k+L)-\boldsymbol{h}^{\mathrm{T}}(k+L)\hat{\boldsymbol{\theta}}(k,k+L-1)),\\ \boldsymbol{K}(k,k+L)=\boldsymbol{P}(k,k+L-1)\boldsymbol{h}(k+L)(1+\boldsymbol{h}^{\mathrm{T}}(k+L)\boldsymbol{P}(k,k+L-1)\boldsymbol{h}(k+L))^{-1},\\ \boldsymbol{P}(k,k+L)=(\boldsymbol{I}-\boldsymbol{K}(k,k+L)\boldsymbol{h}^{\mathrm{T}}(k+L))\boldsymbol{P}(k,k+L-1).\end{cases}$$

上述算法前三个式子用于去掉老数据的信息，后三个式子用来增加新数据的信息，初始值取

$$\begin{cases}\boldsymbol{P}(0,0)=\alpha^2\boldsymbol{I},\\ \hat{\boldsymbol{\theta}}(0,0)=\boldsymbol{\varepsilon},\end{cases}$$

其中 α 为充分大的实数；$\boldsymbol{\varepsilon}$ 为充分小的实向量．相应的准则函数递推计算式为

$$J(k+1,k+L)=J(k,k+L-1)-\frac{\tilde{z}_1^2(k)}{1-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k,k+L)\boldsymbol{h}(k)}+\frac{\tilde{z}_2^2(k+L)}{1+\boldsymbol{h}^{\mathrm{T}}(k+L)\boldsymbol{P}(k,k+L-1)\boldsymbol{h}(k+L)},$$

其中

$$\begin{cases}\tilde{z}_1(k)=z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k,k+L),\\ \tilde{z}_2(k+L)=z(k+L)-\boldsymbol{h}^{\mathrm{T}}(k+L)\hat{\boldsymbol{\theta}}(k,k+L-1).\end{cases}$$

3.8.4　折息法

折息法把加权最小二乘法和遗忘因子法融合起来，形成如下递推算法：

$$\begin{cases}\hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\boldsymbol{K}(k)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)),\\ \boldsymbol{K}(k)=\boldsymbol{P}(k-1)\boldsymbol{h}(k)\left(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+\dfrac{\mu(k)}{\Lambda(k)}\right)^{-1},\\ \boldsymbol{P}(k)=\dfrac{1}{\mu(k)}(\boldsymbol{I}-\boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1).\end{cases}$$

折息因子与加权因子和遗忘因子之间的关系为 $\Gamma(k,i)=\Lambda(i)\prod_{j=i+1}^{k}\mu(j)$，当遗忘因子取常数时，折息因子又可表示成 $\Gamma(k,i)=\Lambda(i)\mu^{k-i}$．折息法同时具有加权最小二乘法和遗忘因子法的作用，既可获得系统的平均特性，又具有时变跟踪能力．

3.8.5 协方差重调最小二乘算法

在辨识递推计算过程中，协方差矩阵 $\boldsymbol{P}(k)$ 衰减很快，此时算法的增益矩阵 $\boldsymbol{K}(k)$ 也急剧衰减．这种现象的出现，促使人们去考虑一种修正的方案，即在指定的时刻重新调整协方差矩阵 $\boldsymbol{P}(k)$，使算法始终保持较快的收敛速度．这种协方差重调的最小二乘算法可描述如下：

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+1)^{-1}, \\ \boldsymbol{P}(k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1).\end{cases}$$

当 $k \notin \{k_1, k_2, \cdots, k_l\}$ 时，$\boldsymbol{P}(k)$ 按上式算法计算；当 $k = k_i \in \{k_1, k_2, \cdots, k_l\}$ 时，把 $\boldsymbol{P}(k)$ 重调为

$$\boldsymbol{P}(k_i) = \alpha_i \boldsymbol{I}, \qquad 0 < \alpha_{\min} \leqslant \alpha_i \leqslant \alpha_{\max} < \infty.$$

3.8.6 协方差修正最小二乘算法

对时变系统辨识来说，为了防止矩阵 $\boldsymbol{P}(k)$ 趋于零，当参数估计值超过某阈值时，矩阵 $\boldsymbol{P}(k)$ 将自动加上附加项 $\boldsymbol{Q}$，具体算法如下：

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+1)^{-1}, \\ \overline{\boldsymbol{P}}(k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1), \\ \boldsymbol{P}(k) = \overline{\boldsymbol{P}}(k) + \boldsymbol{Q} \qquad (\boldsymbol{Q} \geqslant 0).\end{cases}$$

3.8.7 有约束的最小二乘算法

在辨识问题中，有可能需要把参数估计值约束在某个范围内．设要考虑的约束范围是参数空间中的一个闭凸集，记作 $\mathscr{B}$. 如果参数估计值 $\hat{\boldsymbol{\theta}}(k) \in \mathscr{B}$，则按最小二乘算法(3-16) 式继续计算；否则需进行下列变换：

(1) 建立参数空间的变换坐标基

$$\boldsymbol{\rho} = \boldsymbol{P}^{-\frac{1}{2}}(k)\boldsymbol{\theta},$$

其中 $(\boldsymbol{P}^{-1/2}(k))^{\mathrm{T}}\boldsymbol{P}^{-1/2}(k) = \boldsymbol{P}^{-1}(k)$.

(2) 在变换坐标 $\boldsymbol{P}^{-\frac{1}{2}}(k)$ 下，将参数约束空间 $\mathscr{B}$ 变换成像空间 $\bar{\mathscr{B}}$.

(3) 在变换坐标 $\boldsymbol{P}^{-\frac{1}{2}}(k)$ 下，将参数估计 $\hat{\boldsymbol{\theta}}(k)$ 变换成像 $\hat{\boldsymbol{\rho}}(k)$，即作如下变换：

$$\hat{\boldsymbol{\rho}}(k) = \boldsymbol{P}^{-\frac{1}{2}}(k)\hat{\boldsymbol{\theta}}(k).$$

(4) 将参数估计值 $\hat{\boldsymbol{\theta}}(k)$ 的像 $\hat{\boldsymbol{\rho}}(k)$ 正交投影到像空间 $\bar{\mathscr{B}}$ 的边界上，生成边界上的像 $\hat{\boldsymbol{\rho}}^{\mathrm{b}}(k)$.

(5) 在变换坐标 $\boldsymbol{P}^{\frac{1}{2}}(k)$ 下，将像空间 $\bar{\mathscr{B}}$ 的边界上的像 $\hat{\boldsymbol{\rho}}^{\mathrm{b}}(k)$ 变换回参数估计

值 $\hat{\boldsymbol{\theta}}(k)$，即作如下变换：

$$\hat{\boldsymbol{\theta}}(k) = \boldsymbol{P}^{\frac{1}{2}}(k)\hat{\boldsymbol{\rho}}^{b}(k).$$

这时参数估计值 $\hat{\boldsymbol{\theta}}(k)$ 又将返回到约束空间 $\mathscr{B}$ 内，然后继续进行运算．

3.9 最小二乘算法的UD分解实现

考虑模型(3-3)式最小二乘参数辨识问题，设噪声 $n(k)$ 为零均值白噪声，记作 $v(k)$，又设(3-4)式迟延算子多项式 $A(Z^{-1})$ 和 $B(Z^{-1})$ 的阶次相等，取作 n，则模型(3-3)式可成如下的最小二乘格式：

$$z(k) = \boldsymbol{h}_n^{\mathrm{T}}(k)\boldsymbol{\theta}_n + v(k),$$

其中

$$\begin{cases}\boldsymbol{\theta}_n = (a_n, b_n, a_{n-1}, b_{n-1}, \cdots, a_1, b_1)^{\mathrm{T}},\\ \boldsymbol{h}_n(k) = (-z(k-n), u(k-n), \cdots, -z(k-1), u(k-1))^{\mathrm{T}}.\end{cases}$$

又定义

$$\boldsymbol{\varphi}_n(k) = (-z(k-n), u(k-n), \cdots, -z(k-1), u(k-1), -z(k))^{\mathrm{T}},$$

使数据向量 $\boldsymbol{h}_n(k)$ 和 $\boldsymbol{\varphi}_n(k)$ 具有“移位”性质，即

$$\boldsymbol{\varphi}_n(k) = \begin{bmatrix}\boldsymbol{h}_n(k)\\ -z(k)\end{bmatrix},$$

$$\boldsymbol{h}_n(k) = \begin{bmatrix}\boldsymbol{\varphi}_{n-1}(k-1)\\ u(k-1)\end{bmatrix}.$$

置 $$\boldsymbol{R}_n(k) = \sum_{j=1}^{k}\boldsymbol{h}_n(j)\boldsymbol{h}_n^{\mathrm{T}}(j) \text{ 和 } \boldsymbol{S}_n(k) = \sum_{j=1}^{k}\boldsymbol{\varphi}_n(j)\boldsymbol{\varphi}_n^{\mathrm{T}}(j),$$

利用“移位”性质，可将矩阵 $\boldsymbol{S}_n(k)$ 分解成

$$\boldsymbol{S}_n(k) = \begin{bmatrix}\boldsymbol{I}_{2n} & \boldsymbol{0}\\ -\hat{\boldsymbol{\theta}}_n^{\mathrm{T}}(k) & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_n(k) & \boldsymbol{0}\\ \boldsymbol{0} & J_n(k)\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_{2n} & \boldsymbol{0}\\ -\hat{\boldsymbol{\theta}}_n^{\mathrm{T}}(k) & 1\end{bmatrix}^{\mathrm{T}}.$$

其中 $\boldsymbol{I}_{2n}$ 为 $2n$ 维单位阵；$\hat{\boldsymbol{\theta}}_n(k)$ 和 $J_n(k)$ 分别为模型阶次取 n 时的参数估计值和对应的准则函数，其表达式分别为

$$\begin{cases}\hat{\boldsymbol{\theta}}_n(k) = \boldsymbol{R}_n^{-1}(k)\sum_{j=1}^{k}\boldsymbol{h}_n(j)z(j),\\ J_n(k) = \sum_{j=1}^{k}z^2(j) - \hat{\boldsymbol{\theta}}_n^{\mathrm{T}}(k)\boldsymbol{R}_n(k)\hat{\boldsymbol{\theta}}_n(k).\end{cases}$$

同样，矩阵 $\boldsymbol{R}_n(k)$ 亦可分解成

$$\boldsymbol{R}_n(k) = \begin{bmatrix}\boldsymbol{I}_{2n-1} & \boldsymbol{0}\\ -\hat{\boldsymbol{\theta}}_{n-1}^{*\mathrm{T}}(k-1) & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{S}_{n-1}(k-1) & \boldsymbol{0}\\ \boldsymbol{0} & J_{n-1}^{*}(k-1)\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_{2n-1} & \boldsymbol{0}\\ -\hat{\boldsymbol{\theta}}_{n-1}^{*\mathrm{T}}(k-1) & 1\end{bmatrix}^{\mathrm{T}}.$$

其中 $\boldsymbol{I}_{2n-1}$ 为 $2n-1$ 维单位阵；$\hat{\boldsymbol{\theta}}_{n-1}^{*}(k-1)$ 和 $\hat{J}_{n-1}^{*}(k-1)$ 分别为

$$\begin{cases}\hat{\boldsymbol{\theta}}_{n-1}^{*}(k-1)=\boldsymbol{S}_{n-1}^{-1}(k-1)\sum_{j=1}^{k}\boldsymbol{\varphi}_{n}(j)u(j),\\ J_{n-1}^{*}(k-1)=\sum_{j=0}^{k}u^{2}(j)-\hat{\boldsymbol{\theta}}_{n-1}^{*\mathrm{T}}(k-1)\boldsymbol{S}_{n-1}^{-1}(k-1)\hat{\boldsymbol{\theta}}_{n-1}^{*}(k-1).\end{cases}$$

重复 $\boldsymbol{R}_n(k)$ 和 $\boldsymbol{S}_n(k)$ 上述的迭代关系，可得到矩阵 $\boldsymbol{S}_n^{-1}(k)$ 的UD分解形式为

$$\boldsymbol{S}_n^{-1}(k)=\boldsymbol{U}_n(k)\boldsymbol{D}_n(k)\boldsymbol{U}_n^{\mathrm{T}}(k),$$

其中

$$\boldsymbol{U}_n(k)=\begin{bmatrix}1 & \hat{\boldsymbol{\theta}}_0^{*}(k-n) & & & & & \\ & 1 & \hat{\boldsymbol{\theta}}_1(k-n+1) & & & \boldsymbol{0} & \\ & & 1 & \ddots & & & \\ & & & \ddots & \hat{\boldsymbol{\theta}}_{n-1}^{*}(k-1) & & \\ & \boldsymbol{0} & & & 1 & \hat{\boldsymbol{\theta}}_n(k) \\ & & & & & 1\end{bmatrix},$$

$$\boldsymbol{D}_n(k)=\begin{bmatrix}J_0^{-1}(k-n) & & & & & \\ & J_0^{*-1}(k-n) & & & \boldsymbol{0} & \\ & & J_1^{-1}(k-n+1) & & & \\ & & & \ddots & & \\ & \boldsymbol{0} & & & J_{n-1}^{*-1}(k-1) & \\ & & & & & J_n^{-1}(k)\end{bmatrix}.$$

令 $\boldsymbol{C}_n(k)=\boldsymbol{S}_n^{-1}(k)$，由上述讨论表明，$\boldsymbol{C}_n(k)$ 矩阵包含的信息都是有用的，其中 $\hat{\boldsymbol{\theta}}_i(k-n+i)$ 和 $J_i(k-n+i)$，$i=1,2,\cdots,n$，分别代表 $k-n+i$ 时刻 i 阶模型参数估计值和准则函数．也就是说，从1阶到 n 阶的参数估计值和准则函数都浓缩在矩阵 $\boldsymbol{C}_n(k)$ 中．对矩阵 $\boldsymbol{C}_n(k)$ 进行UD分解，即可一次获得各阶次的模型参数估计值和准则函数，以此可以同时确定模型的阶次和参数估计值．下面把最小二乘算法的UD分解递推计算步骤归纳为：

(1) 初始化 $\boldsymbol{U}_n(0)$ 和 $\boldsymbol{D}_n(0)$，使 $\boldsymbol{U}_n(0)\boldsymbol{D}_n(0)\boldsymbol{U}_n^{\mathrm{T}}(0)=\alpha^2\boldsymbol{I}$，其中 α 为充分大的实数．

(2) 由 $\boldsymbol{U}_n(k-1)$ 和 $\boldsymbol{D}_n(k-1)$，计算

$$\begin{cases}\boldsymbol{g}_n(k)=[g_1(k),g_2(k),\cdots,g_N(k)]^{\mathrm{T}}=\boldsymbol{D}_n(k-1)\boldsymbol{f}_n(k),\\ \boldsymbol{f}_n(k)=[f_1(k),f_2(k),\cdots,f_N(k)]^{\mathrm{T}}=\boldsymbol{U}_n^{\mathrm{T}}(k-1)\boldsymbol{\varphi}_n(k),\\ N=\dim\boldsymbol{\theta}+1=2n+1,\end{cases}$$

其中 $\boldsymbol{\varphi}_n(k)$ 为数据向量，是可测的．

(3) 置 $s_0(k)=\mu(k)$,其中 $\mu(k)$ 为引入的遗忘因子;从 $j=1$ 到 N,第(4)~(7)步计算.

(4) 计算
$$\begin{cases} s_j(k)=s_{j-1}(k)+f_j(k)g_j(k), \\ d_j(k)=\dfrac{d_j(k-1)s_{j-1}(k)}{s_j(k)\mu(k)}, \\ \gamma_j(k):=g_j(k), \\ \lambda_j(k):=\dfrac{-f_j(k)}{s_{j-1}(k)}, \end{cases}$$
其中 $d_j(k)$ 代表 $\boldsymbol{D}_n(k)$ 第 j 个对角线元素;$X:=Y$ 表示将 Y 置赋给 X.

(5) 从 $i=1$ 到 $i=j-1$,按第(6)~(7)步计算,若 $j=1$,跳回第(4)步.

(6) 计算
$$\begin{cases} u_{ij}(k)=u_{ij}(k-1)+\gamma_i(k)\lambda_j(k), \\ \gamma_i(k):=\gamma_i(k)+u_{ij}(k-1)\gamma_j(k). \end{cases}$$
其中 $u_{ij}(k)$ 代表 $\boldsymbol{U}_n(k)$ 第 i 行第 j 列元素.

(7) 根据矩阵 $\boldsymbol{U}_n(k)$ 和 $\boldsymbol{D}_n(k)$ 的组成结构,确定各阶模型的参数估计值和准则函数.

(8) 置 $k:=k+1$,返回第(2)步.如此不断递推下去,直至辨识计算结束为止.

4 最小二乘类法

最小二乘法是一种最基本的辨识方法,但如果模型的噪声不是白噪声,用最小二乘法不能得到无偏、一致估计.本章着重讨论当模型噪声为有色噪声时,各种最小二乘辨识的方法.

4.1 增广最小二乘法

4.1.1 增广最小二乘法原理

考虑如下模型:
$$A(Z^{-1})z(k)=B(Z^{-1})u(k)+N(Z^{-1})v(k). \tag{4-1}$$
其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量;$v(k)$ 是均值为零、方差为 σ_v^2 的不相关随机噪声,或称白噪声;$N(Z^{-1})$ 为噪声模型;$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式,记作
$$\begin{cases} A(Z^{-1})=1+a_1Z^{-1}+a_2Z^{-2}+\cdots+a_{n_a}Z^{-n_a}, \\ B(Z^{-1})=b_1Z^{-1}+b_2Z^{-2}+\cdots+b_{n_b}Z^{-n_b}, \end{cases}$$
其中 n_a 和 n_b 为模型阶次.为了运用最小二乘原理来辨识这种模型的参数,需要把模型(4-1)式写成最小二乘格式:
$$z(k)=\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta}+v(k), \tag{4-2}$$
这样就必须把噪声模型的参数包含在参数向量 $\boldsymbol{\theta}$ 中,从而引出"增广"的概念,用

它来构造(4-2)式的参数向量 $\boldsymbol{\theta}$ 和数据向量 $\boldsymbol{h}(k)$,具体的构成形式会因噪声模型的结构不同而不同．下面是三种不同噪声模型的数据向量构成方法.

(1) 若 $N(Z) = D(Z^{-1}) = 1 + d_1Z^{-1} + d_2Z^{-2} + \cdots + d_{n_d}Z^{-n_d}$,则可按下式分别构成参数向量和数据向量:

$$\begin{cases} \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), u(k-1), u(k-2), \\ \qquad \cdots, u(k-n_b), \hat{v}(k-1), \hat{v}(k-2), \cdots, \hat{v}(k-n_d))^{\mathrm{T}}, \\ \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}, \\ \hat{v}(k) = z(k) + \sum_{i=1}^{n_a} \hat{a}_i(k-1)z(k-i) - \sum_{i=1}^{n_b} \hat{b}_i(k-1)u(k-i) - \sum_{i=1}^{n_d} \hat{d}_i(k-1)\hat{v}(k-i). \end{cases}$$

(2) 若 $N(z) = \dfrac{1}{C(z^{-1})} = \dfrac{1}{1 + c_1z^{-1} + c_2z^{-2} + \cdots + c_{n_c}z^{-n_c}}$,则参数向量和数据向量分别为

$$\begin{cases} \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), u(k-1), u(k-2), \\ \qquad \cdots, u(k-n_b), -\hat{e}(k-1), -\hat{e}(k-2), \cdots, -\hat{e}(k-n_c))^{\mathrm{T}}, \\ \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c})^{\mathrm{T}}, \\ \hat{E}(k) = z(k) + \sum_{i=1}^{n_a} \hat{a}_i(k-1)z(k-i) - \sum_{i=1}^{n_b} \hat{b}_i(k-1)u(k-i). \end{cases}$$

(3) 若 $N(z) = \dfrac{D(z^{-1})}{C(z^{-1})} = \dfrac{1 + d_1z^{-1} + d_2z^{-2} + \cdots + d_{n_d}z^{-n_d}}{1 + c_1z^{-1} + c_2z^{-2} + \cdots + c_{n_c}z^{-n_c}}$,

则参数向量和数据向量分别为

$$\begin{cases} \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), u(k-1), u(k-2), \cdots, u(k-n_b), \\ \qquad -\hat{e}(k-1), -\hat{e}(k-2), \cdots, -\hat{e}(k-n_c), \hat{v}(k-1), \hat{v}(k-2), \cdots, \hat{v}(k-n_d))^{\mathrm{T}}, \\ \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c}, d_1, d_2, \cdots, d_{n_c})^{\mathrm{T}}, \\ \hat{v}(k) = \hat{e}(k) + \sum_{i=1}^{n_c} \hat{c}_i(k-1)\hat{e}(k-i) - \sum_{i=1}^{n_d} \hat{d}_i(k-1)\hat{v}(k-i), \\ \hat{e}(k) = z(k) + \sum_{i=1}^{n_a} \hat{a}_i(k-1)z(k-i) - \sum_{i=1}^{n_b} \hat{b}_i(k-1)u(k-i). \end{cases}$$

以上这种构成参数向量和数据向量的思想就是所谓的增广原理,它是增广最小二乘法的根本．

4.1.2 增广最小二乘算法

对模型(4-2)式运用最小二乘原理,可导出如下的增广最小二乘批处理算法:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ELS}} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{z}_L,$$

其中

$$\begin{cases} z_L = (z(1), z(2), \cdots, z(L))^{\mathrm{T}}, \\ \boldsymbol{H}_L = \begin{bmatrix} \boldsymbol{h}^{\mathrm{T}}(1) \\ \boldsymbol{h}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}^{\mathrm{T}}(L) \end{bmatrix}. \end{cases}$$

又有如下与之对应的参数估计递推算法，记作 **RELS**(recursive extended least squares algorithm)：

$$\begin{cases} \hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k) + 1)^{-1}, \\ \boldsymbol{P}(k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1), \end{cases}$$

其中数据向量 $\boldsymbol{h}(k)$ 视不同的噪声模型可以具有不同的增广结构．

这种参数估计算法的实质，在于把噪声模型参数混在参数向量 $\boldsymbol{\theta}$ 中一起进行辨识．就这种意义上说，可称之为增广最小二乘法．它是普通最小二乘法的一种推广，其递推形式和性质与普通最小二乘法的完全相同．优点是可用来解决有色噪声模型的辨识问题，但噪声模型部分本身的辨识并不是无偏、一致估计．

4.1.3　增广最小二乘算法的 UD 分解实现

当(4-1)式噪声模型取 $N(Z) = D(Z^{-1})$，又设迟延算子多项式 $A(Z^{-1})$、$B(Z^{-1})$ 和 $D(z^{-1})$ 的阶次相等，都取作 n 时，模型(4-1)式的最小二乘格式为

$$z(k) = \boldsymbol{h}_n^{\mathrm{T}}(k)\boldsymbol{\theta}_n + v(k),$$

其中

$$\begin{cases} \boldsymbol{\theta}_n = (a_n, d_n, b_n, a_{n-1}, d_{n-1}, b_{n-1}, \cdots, a_1, d_1, b_1)^{\mathrm{T}}; \\ \boldsymbol{h}_n(k) = (-z(k-n), \hat{v}(k-n), u(k-n), -z(k-(n-1)), \hat{v}(k-(n-1)), \\ \qquad u(k-(n-1)), \cdots, -z(k-1), \hat{v}(k-1), u(k-1))^{\mathrm{T}}; \\ \hat{v}(k) = z(k) - \boldsymbol{h}_n^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}_n(k-1). \end{cases}$$

若在数据向量中加入当前的数据信息，则可构成具有"移位"性质的新数据向量为

$$\boldsymbol{\phi}_n(k) = \begin{bmatrix} \boldsymbol{h}_n(k) \\ -z(k) \end{bmatrix}, \boldsymbol{\varphi}_n(k) = \begin{bmatrix} \boldsymbol{\phi}_n(k) \\ \hat{v}(k) \end{bmatrix},$$

并有

$$\boldsymbol{h}_n(k) = \begin{bmatrix} \boldsymbol{\varphi}_{n-1}(k-1) \\ u(k-1) \end{bmatrix}.$$

上述这三个数据向量构成的"移位"结构是利用 UD 分解方法来实现增广最小二乘算法的基础．和 3.9 节最小二乘算法的 UD 分解实现的步骤一样，对矩阵 $\boldsymbol{C}_n(k) = \boldsymbol{S}_n^{-1}(k) = \left[\sum_{j=1}^{k}\boldsymbol{\phi}_n(j)\boldsymbol{\phi}_n^{\mathrm{T}}(j)\right]^{-1}$ 进行 UD 分解，可同时获得各阶模型的参

数估计值和准则函数 .

4.2　辅助变量法

4.2.1　辅助变量法原理

考虑如下模型:

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + e(k), \tag{4-3}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量;$e(k)$ 是均值为零、方差为 σ_e^2 的有色噪声;$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式,分别记作

$$\begin{cases} A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_{n_a}Z^{-n_a}, \\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_{n_b}Z^{-n_b}, \end{cases}$$

其中 n_a 和 n_b 为模型阶次 . 由于 $e(k)$ 是有色噪声,利用最小二乘原理不能获得模型参数的无偏、有效、一致估计 .

把模型(4-3) 式写成最小二乘格式:

$$\boldsymbol{z}_L = \boldsymbol{H}_L\boldsymbol{\theta} + \boldsymbol{e}_L, \tag{4-4}$$

其中

$$\begin{cases} \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{\mathrm{T}}, \\ \boldsymbol{z}_L = (z(1), z(2), \cdots, z(L))^{\mathrm{T}}; \\ \boldsymbol{H}_L = \begin{bmatrix} \boldsymbol{h}^{\mathrm{T}}(1) \\ \boldsymbol{h}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}^{\mathrm{T}}(L) \end{bmatrix}, \\ \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), \\ \qquad u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}; \\ \boldsymbol{e}_L = (e(1), e(2), \cdots, e(L))^{\mathrm{T}}. \end{cases}$$

为了获得(4-4) 式模型参数的无偏、一致估计,准则函数不能再用(3-9) 式的形式 . 设模型残差为

$$\boldsymbol{\varepsilon}_L = \boldsymbol{z}_L - \boldsymbol{H}_L\boldsymbol{\theta}.$$

可以证明,如果能够设法使上式中的残差序列 $\boldsymbol{\varepsilon}_L = (\varepsilon(1), \varepsilon(2), \cdots, \varepsilon(L))^{\mathrm{T}}$ 与过去的数据序列不相关,则可获得模型参数 $\boldsymbol{\theta}$ 的无偏估计 . 为此,就须从过去的数据集合中设法衍生出一个有限维的向量 $\boldsymbol{h}^*(k)$,使之与残差序列不相关,即

$$\lim_{L\to\infty}\sum_{k=1}^{L}\boldsymbol{h}^*(k)\varepsilon(k) = 0.$$

这样,准则函数就可写成如下形式:

$$J(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}_L^{\mathrm{T}}\boldsymbol{H}_L^*\boldsymbol{H}_L^{*\mathrm{T}}\boldsymbol{\varepsilon}_L, \tag{4-5}$$

其中

$$H_L^* = \begin{bmatrix} h^{*\mathrm{T}}(1) \\ h^{*\mathrm{T}}(2) \\ \vdots \\ h^{*\mathrm{T}}(L) \end{bmatrix}.$$

比较(4-5)式与(3-9)式可知,(4-5)式可以看做加权阵取 $\boldsymbol{\Lambda}_L = H_L^* H_L^{*\mathrm{T}}$ 的一种加权准则函数．对于极小化(4-5)式,当 $H_L^{*\mathrm{T}} H_L^*$ 非奇异时,可以获得(4-4)式模型参数的辅助变量估计

$$\hat{\theta}_{\mathrm{IV}} = (H_L^{*\mathrm{T}} H_L)^{-1} H_L^{*\mathrm{T}} z_L. \tag{4-6}$$

这种辨识思想称为辅助变量原理,对应的辨识方法称为辅助变量法．$h^*(k)$ 是从过去数据集合中引申出来的向量,称为辅助向量．由此组成的 H_L^* 称为辅助矩阵．不难证明,如果下面两个条件能够满足,即

1° $\frac{1}{L} H_L^{*\mathrm{T}} H_L \xrightarrow{k \to \infty} E\{h^*(k) h(k)\}$ 是非奇异的,

2° $\frac{1}{L} H_L^{*\mathrm{T}} e_L \xrightarrow{k \to \infty} E\{h^*(k) e(k)\} = 0$, (W.P.1.),

则由算法(4-6)式获得的模型参数估计是收敛的,即有

$$\hat{\theta}_{\mathrm{IV}} \xrightarrow{L \to \infty} \theta_0 \quad (\mathrm{W.P.1.}).$$

通过适当选择辅助变量向量 $h^*(k)$,上述两个条件是可以满足的．

4.2.2　辅助向量的选择

当噪声 $e(k)$ 与模型输入 $u(k)$ 不相关,且输入 $u(k)$ 为持续激励信号时,可按下列方法之一选择辅助向量:

(1)
$$\begin{cases} h^*(k) = (-x(k-1), -x(k-2), \cdots, -x(k-n_a), \\ \qquad\qquad u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}, \\ x(k) = h^{*\mathrm{T}}(k) \bar{\theta}(k), \\ \bar{\theta}(k) = (1-a)\bar{\theta}(k-1) + a\hat{\theta}(k-d) \quad (a = 0.01 \sim 0.1, d = 0 \sim 10), \end{cases}$$

(2) $h^*(k) = (-u(k-n_b-1), -u(k-n_b-2), \cdots, -u(k-n_b-n_a), u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}$,

(3) $h^*(k) = (-z(k-n_d-1), -z(k-n_d-2), \cdots, -z(k-n_d-n_a), u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}$,

(4) $h^*(k) = (u(k-1), u(k-2), \cdots, u(k-n_a-n_b))^{\mathrm{T}}$,

其中 n_d 为噪声模型 $e(k) = D(Z^{-1}) v(k)$ 的阶次．

特别应该指出,如果选第(4)种辅助向量,以此构成的辅助变量法相当于另一种称作**相关二步法**的辨识方法．这可简要分析如下．

(1) 当辅助向量选

$$h^*(k) = (u(k-1), u(k-2), \cdots, u(k-n_a-n_b))^{\mathrm{T}}$$

时,所构成的辅助变量算法相当是如下正则方程:

$$\sum_{k=1}^{L} \boldsymbol{h}^{*}(k)z(k) = \left(\sum_{k=1}^{L} \boldsymbol{h}^{*}(k)\boldsymbol{h}^{\mathrm{T}}(k)\right)\boldsymbol{\theta} + \boldsymbol{v}_{L}.$$

的最小二乘解,其中 $\boldsymbol{v}_L = (v(1),v(2),\cdots,v(L))^{\mathrm{T}}$,为模型噪声向量.

(2) 如果数据是平稳的,那么上述正则方程也可写成以相关函数表示的最小二乘格式:

$$\begin{cases} R_{uz}(l \mid k) = \boldsymbol{h}^{\mathrm{T}}(l \mid k)\boldsymbol{\theta} + v(k), \\ \boldsymbol{h}(l \mid k) = (-R_{uz}(l-1 \mid k),\cdots,-R_{uz}(l-n_a \mid k),R_u(l-1 \mid k),\cdots,R_u(l-n_b \mid k))^{\mathrm{T}}, \end{cases}$$

其中相关函数定义为

$$\begin{cases} R_{uz}(l \mid k) = \dfrac{1}{L}\displaystyle\sum_{k=1}^{L} u(k-l)z(k), \\ R_u(l \mid k) = \dfrac{1}{L}\displaystyle\sum_{k=1}^{L} u(k-l)u(k). \end{cases}$$

(3) 对上述以相关函数表示的正则方程,运用最小二乘原理,亦可获得模型(4-4)式的参数估计值.

(4) 这种先建立相关函数正则方程,即模型(4-4)式两边同乘以 $u(k-l)$,并取数学期望,再求最小二乘解的方法,就称为相关二步法.它与辅助向量取 $\boldsymbol{h}^{*}(k) = (u(k-1),\cdots,u(k-n_a-n_b))^{\mathrm{T}}$ 时的辅助变量算法是等价的.

4.2.3 递推算法

和推导最小二乘递推算法一样,由辅助变量批处理算法(4-6)式可导出如下的递推计算形式,记作 **RIV**(recursive instrumental variable algorithm):

$$\begin{cases} \hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}^{*}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}^{*}(k)+1)^{-1}, \\ \boldsymbol{P}(k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1). \end{cases}$$

同理,也可推导出辅助变量法的残差 $\varepsilon(k)$ 与新息 $\tilde{z}(k)$ 的关系

$$\varepsilon(k) = \frac{\tilde{z}(k)}{1 + \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}^{*}(k)},$$

或

$$\varepsilon(k) = [1 - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{h}^{*}(k)]\tilde{z}(k).$$

以及准则函数 $J(k)$ 的递推计算式

$$J(k) = J(k-1) + \frac{\tilde{z}^{2}(k)}{\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}^{*}(k)+1},$$

其中 $\tilde{z}(k) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)$,是 k 时刻的新息.

辅助变量法的思想是可贵的,适当选择辅助向量就可以沟通它与其他各种辨识算法之间的联系.辅助变量法可以适用于有色噪声模型的辨识,但对参数估计的初始值比较敏感.

4.3　广义最小二乘法

4.3.1　批处理算法

考虑如下模型：

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + \frac{1}{C(Z^{-1})}v(k). \tag{4-7}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$v(k)$ 是均值为零、方差为 σ_v^2 的白噪声；迟延算子多项式 $A(Z^{-1})$、$B(Z^{-1})$ 和 $C(Z^{-1})$ 分别记作

$$\begin{cases} A(Z^{-1}) = 1 + a_1 Z^{-1} + a_2 Z^{-2} + \cdots + a_{n_a} Z^{-n_a}, \\ B(Z^{-1}) = b_1 Z^{-1} + b_2 Z^{-2} + \cdots + b_{n_b} Z^{-n_b}, \\ C(Z^{-1}) = 1 + c_1 Z^{-1} + c_2 Z^{-2} + \cdots + c_{n_c} Z^{-n_c}, \end{cases}$$

其中 n_a、n_b 和 n_c 为模型阶次．令

$$e(k) = \frac{1}{C(Z^{-1})}v(k),$$

且定义

$$\begin{cases} \boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{\mathrm{T}}, \\ \boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), \\ \qquad u(k-1), u(k-2), \cdots, u(k-n_b))^{\mathrm{T}}, \end{cases}$$

则模型(4-7)式可写成如下的最小二乘格式：

$$\boldsymbol{z}_L = \boldsymbol{H}_L\boldsymbol{\theta} + \boldsymbol{e}_L, \tag{4-8}$$

其中

$$\begin{cases} \boldsymbol{z}_L = (z(1), z(2), \cdots, z(L))^{\mathrm{T}}, \\ \boldsymbol{H}_L = \begin{bmatrix} \boldsymbol{h}^{\mathrm{T}}(1) \\ \boldsymbol{h}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}^{\mathrm{T}}(L) \end{bmatrix}, \\ \boldsymbol{e}_L = (e(1), e(2), \cdots, e(L))^{\mathrm{T}}. \end{cases}$$

为了获得(4-8)式模型参数的无偏、一致估计，可运用马尔可夫估计算法，获得如下的模型参数估计

$$\hat{\boldsymbol{\theta}}_{\mathrm{GLS}} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_e^{-1}\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{\Sigma}_e^{-1}\boldsymbol{z}_L, \tag{4-9}$$

其中 $\boldsymbol{\Sigma}_e$ 为噪声向量 $\boldsymbol{e}_L$ 的协方差阵．根据噪声 $e(k)$ 和 $v(k)$ 之间的关系，可求得

$$\boldsymbol{\Sigma}_e = \sigma_v^2(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1},$$

其中

$$C = \begin{bmatrix} 1 & & & & & & \\ c_1 & 1 & & & 0 & & \\ \vdots & c_1 & 1 & & & & \\ c_{n_c} & \vdots & c_1 & 1 & & & \\ 0 & c_{n_c} & \vdots & c_1 & 1 & & \\ \vdots & \ddots & \ddots & \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & c_{n_c} & \cdots & c_1 & 1 \end{bmatrix}.$$

若令 $\boldsymbol{H}_f = \boldsymbol{C}\boldsymbol{H}_L, \boldsymbol{z}_f = \boldsymbol{C}\boldsymbol{z}_L$,则算法(4-9)式可写成

$$\hat{\boldsymbol{\theta}}_{\mathrm{GLS}} = (\boldsymbol{H}_f^{\mathrm{T}}\boldsymbol{H}_f)^{-1}\boldsymbol{H}_f^{\mathrm{T}}\boldsymbol{z}_f. \tag{4-10}$$

如果噪声模型已知,则由(4-10)式可直接估计出模型参数 $\boldsymbol{\theta}$. 如果噪声模型未知,则要用迭代的方法先求得参数 $\boldsymbol{\theta}$ 的估计值,再求噪声模型参数 $\boldsymbol{\theta}_e = (c_1, c_2, \cdots, c_{n_c})^{\mathrm{T}}$ 的估计值. 求 $\boldsymbol{\theta}_e$ 估计值的方法依然可以是最小二乘法. 把噪声 $e(k)$ 和 $v(k)$ 之间的关系写成最小二乘格式,有

$$e(k) = \boldsymbol{h}_e^{\mathrm{T}}(k)\boldsymbol{\theta}_e + v(k),$$

其中

$$\boldsymbol{h}_e(k) = (e(k-1), e(k-2), \cdots, e(k-n_c))^{\mathrm{T}},$$

则有

$$\hat{\boldsymbol{\theta}}_e = (\boldsymbol{H}_e^{\mathrm{T}}\boldsymbol{H}_e)^{-1}\boldsymbol{H}_e^{\mathrm{T}}\boldsymbol{z}_e, \tag{4-11}$$

其中

$$\begin{cases} \boldsymbol{z}_e = (z(1), z(2), \cdots, z(L))^{\mathrm{T}}, \\ \boldsymbol{H}_e = \begin{bmatrix} \boldsymbol{h}_e^{\mathrm{T}}(1) \\ \boldsymbol{h}_e^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}_e^{\mathrm{T}}(L) \end{bmatrix}. \end{cases}$$

(4-10)式和(4-11)式构成迭代计算,这种迭代估计模型参数的方法,称为广义最小二乘法. 这种方法的基本思想是对数据先进行一次滤波预处理,然后利用最小二乘算法来辨识模型的参数. 广义最小二乘参数估计存在全局收敛问题,当噪声较大时,一般难以获得好的辨识效果.

4.3.2 递推算法

由(4-10)式和(4-11)式可以进一步推出广义最小二乘递推算法,记作 **RGLS**(recursive generalized least squares algorithm):

$$\begin{cases}\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}_f(k)(z_f(k) - \boldsymbol{h}_f^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)),\\ \boldsymbol{K}_f(k) = \boldsymbol{P}_f(k-1)\boldsymbol{h}_f(k)(\boldsymbol{h}_f^{\mathrm{T}}(k)\boldsymbol{P}_f(k-1)\boldsymbol{h}_f(k)+1)^{-1},\\ \boldsymbol{P}_f(k) = (\boldsymbol{I} - \boldsymbol{K}_f(k)\boldsymbol{h}_f^{\mathrm{T}}(k))\boldsymbol{P}_f(k-1),\\ \hat{\boldsymbol{\theta}}_e(k) = \hat{\boldsymbol{\theta}}_e(k-1) + \boldsymbol{K}_e(k)(\hat{e}(k) - \boldsymbol{h}_e^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}_e(k-1)),\\ \boldsymbol{K}_e(k) = \boldsymbol{P}_e(k-1)\boldsymbol{h}_e(k)(\boldsymbol{h}_e^{\mathrm{T}}(k)\boldsymbol{P}_e(k-1)\boldsymbol{h}_e(k)+1)^{-1},\\ \boldsymbol{P}_e(k) = (\boldsymbol{I} - \boldsymbol{K}_e(k)\boldsymbol{h}_e^{\mathrm{T}}(k))\boldsymbol{P}_e(k-1),\end{cases}$$

其中

$$\begin{cases}\boldsymbol{h}_f(k) = (-z_f(k-1), -z_f(k-2), \cdots, -z_f(k-n_a),\\ \qquad\qquad u_f(k-1), u_f(k-2), \cdots, u_f(k-n_b))^{\mathrm{T}};\\ \boldsymbol{h}_e(k) = (-\hat{e}(k-1), -\hat{e}(k-2), \cdots, -\hat{e}(k-n_c))^{\mathrm{T}},\\ \hat{e}(k) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k).\end{cases}$$

5 梯度校正法

梯度校正参数辨识方法,是沿着准则函数负梯度方向去寻找参数估计值的方法.

5.1 辨识问题及随机逼近解

考虑如下模型辨识问题:

$$z(k) = \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta} + e(k). \tag{5-1}$$

其中 $z(k)$ 为模型输出;$\boldsymbol{h}(k)$ 为数据向量;$\boldsymbol{\theta}$ 为模型参数;$e(k)$ 为零均值噪声. 定义准则函数为

$$J(\boldsymbol{\theta}) = \frac{1}{2}E\{(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta})^2\}, \tag{5-2}$$

通过极小化 $J(\boldsymbol{\theta})$,也就是解下列方程

$$E\{\boldsymbol{h}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta})\} = \boldsymbol{0}, \tag{5-3}$$

可以求得模型参数 $\boldsymbol{\theta}$ 的估计值.

由于噪声 $e(k)$ 的统计性质未知,方程(5-3)式难以求解,若数学期望 $E\{\cdot\}$ 用 $\frac{1}{L}\sum_{k=1}^{L}(\cdot)$ 来近似,则上述辨识问题退化成最小二乘辨识.

依据梯度校正法的思想,使准则函数(5-2)式达到极小值的模型参数估计,可沿着准则函数的负梯度方向搜索,即可将模型参数估计写成

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) - \rho(k)\left(\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\Bigg|_{\hat{\boldsymbol{\theta}}(k-1)}, \tag{5-4}$$

其中负号表示搜索方向．根据(5-2)式定义，准则函数关于参数的一阶梯度为

$$\left(\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} = -E\{\boldsymbol{h}(k)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{\theta})\},$$

利用随机逼近原理，可将(5-4)式写成

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \rho(k)\boldsymbol{h}(k)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)), \tag{5-5}$$

其中 $\rho(k)$ 为随机收敛因子，它的选择应使得

$$J(\boldsymbol{\theta})\mid_{\hat{\boldsymbol{\theta}}(k)} < J(\boldsymbol{\theta})\mid_{\hat{\boldsymbol{\theta}}(k-1)}. \tag{5-6}$$

通过上述迭代算法，可逐步逼近求得方程(5-3)式的解．若选择 $\rho(k)$ 满足下列条件：

1° $\rho(k) > 0 \qquad (\forall k)$;

2° $\lim\limits_{k\to\infty}\rho(k) = 0$;

3° $\sum\limits_{k=1}^{\infty}\rho(k) = \infty$;

4° $\sum\limits_{k=1}^{\infty}\rho^2(k) < \infty$.

则可保证(5-6)式成立，且使模型参数在均方意义下收敛，也就是

$$\lim_{k\to\infty} E\{(\hat{\boldsymbol{\theta}}(k)-\boldsymbol{\theta}_0)^2\} = \boldsymbol{0},$$

其中 θ_0 为模型参数真值．

当选择 $\rho(k) = 1/k$ 时，(5-5)式与(3-15)式的最小二乘递推算法 $\boldsymbol{R}(k)\Lambda(k) = \boldsymbol{I}$ 的情况是等价的．

5.2 随机牛顿辨识算法

算法(5-5)式是梯度校正法思想的具体表现，也称最速下降法．当准则函数值接近极小值时，这种算法的效率变得很低．为提高算法的效率，典型的做法是用随机牛顿方向来代替(5-4)式中的搜索方向．对准则函数(5-2)式来说，随机牛顿方向为

$$-\left(\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)^{-1}\left(\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}.$$

也就是说，这时应沿着随机牛顿方向去搜索模型参数估计值，以使准则函数(5-2)式达到极小值，即

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) - \rho(k)\left(\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)^{-1}\left(\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\Bigg|_{\hat{\boldsymbol{\theta}}(k-1)}. \tag{5-7}$$

令 $\boldsymbol{R}(k) = \dfrac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$，称 $\boldsymbol{R}(k)$ 为黑塞(Hessian)矩阵．依据(5-2)式定义，矩阵 $\boldsymbol{R}(k)$ 可写成

$$\boldsymbol{R}(k) = E\{\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k)\},$$

或

$$E\{\boldsymbol{R}(k) - \boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k)\} = \boldsymbol{0}.$$

再次利用随机逼近原理,可得到黑塞矩阵 $\boldsymbol{R}(k)$ 的递推计算形式:

$$\boldsymbol{R}(k)=\boldsymbol{R}(k-1)+\rho(k)(\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k)-\boldsymbol{R}(k-1)), \tag{5-8}$$

综合(5-7)式和(5-8)式,就可构成(5-1)式模型参数的**随机牛顿辨识算法**:

$$\begin{cases}\hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\rho(k)\boldsymbol{R}(k)\boldsymbol{h}(k)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1));\\ \boldsymbol{R}(k)=\boldsymbol{R}(k-1)+\rho(k)(\boldsymbol{h}(k)\boldsymbol{h}^{\mathrm{T}}(k)-\boldsymbol{R}(k-1)),\end{cases} \tag{5-9}$$

其中 $\rho(k)$ 为随机收敛因子. 当取 $\rho(k)=\dfrac{1}{k}$ 时,算法(5-9)式就是最小二乘递推算法.

5.3 梯度校正辨识算法

令 $\boldsymbol{P}(k)=\rho(k)\boldsymbol{R}^{-1}(k)$,则(5-9)式就成为(5-1)式模型参数的递推辨识算法:

$$\begin{cases}\hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\boldsymbol{K}(k)(z(k)-\boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)),\\ \boldsymbol{K}(k)=\boldsymbol{P}(k-1)\boldsymbol{h}(k)(\boldsymbol{h}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}(k)+\mu(k))^{-1},\\ \boldsymbol{P}(k)=\dfrac{1}{\mu(k)}(\boldsymbol{I}-\boldsymbol{K}(k)\boldsymbol{h}^{\mathrm{T}}(k))\boldsymbol{P}(k-1),\\ \mu(k)=\dfrac{\rho(k-1)}{\rho(k)}(1-\rho(k)),\end{cases}$$

其中 $\mu(k)$ 为遗忘因子. 若遗忘因子为常数,则收敛因子与遗忘因子的关系为

$$\rho(k)=\frac{1-\mu}{1-\mu^{k}}.$$

6 极大似然法

极大似然法是一种参数辨识方法,它是按系统输出的条件概率密度函数最大逼近原则去寻找模型的参数估计值.

6.1 极大似然原理

系统辨识和参数估计涉及到从随机观测数据中提取信息的问题. 如果寻找到模型参数 $\hat{\boldsymbol{\theta}}$,使系统的输出在 $\hat{\boldsymbol{\theta}}$ 条件下的概率密度最大可能地逼近输出在 $\boldsymbol{\theta}_0$ 条件下的概率密度,则认为模型最大限量地包含了数据中的信息. 这种辨识思想可表达成 $p(z(k)\mid\hat{\boldsymbol{\theta}})\xrightarrow{\max}p(z(k)\mid\boldsymbol{\theta}_0)$.

定义1 设系统输出变量 $z(k)$ 在模型参数 $\boldsymbol{\theta}$ 条件下的概率密度为 $p(z(k)\mid\boldsymbol{\theta})$,独立观测的随机数据向量 $\boldsymbol{z}=(z(1),z(2),\cdots,z(L))^{\mathrm{T}}$ 的联合概率密度记作 $p(\boldsymbol{z}\mid\boldsymbol{\theta})$,则对一批确定的观测数据来说,$p(\boldsymbol{z}\mid\boldsymbol{\theta})$ 是以模型参数 $\boldsymbol{\theta}$ 为自变量的概率密度函数,这个函数称为**似然函数**,记作 $L(\boldsymbol{z}\mid\boldsymbol{\theta})$.

$L(z \mid \theta)$ 反映被观测事件确实应当发生的“可能性”，与母集概率密度函数的关系为 $L(z \mid \theta) = \prod_{k=1}^{L} p(z(k) \mid \theta)$，对应的对数似然函数记作 $l(z \mid \theta) = \sum_{k=1}^{L} \ln p(z(k) \mid \theta)$.

定义 2 称
$$I(\theta_0,\theta) = E\{\ln p(z(k) \mid \theta_0)\} - E\{\ln p(z(k) \mid \theta)\} = E\left\{\ln \frac{p(z(k) \mid \theta_0)}{p(z(k) \mid \theta)}\right\}.$$
为库尔贝克 - 莱布勒(Kullback-leibler)信息测度.

可以证明 $I(\theta_0,\theta) \geqslant 0$，可见模型参数 θ 的合理估计是选择它使 $I(\theta_0,\theta)$ 为最小，也就是 $\hat{\theta}_{ML} = \arg\max_{\theta} L(z \mid \theta)$，即 $\hat{\theta}_{ML}$ 使极大似然函数 $L(z \mid \theta)$ 达到最大值. 这时，
$$\left(\frac{\partial L(z \mid \theta)}{\partial \theta}\right)^{T}\Bigg|_{\hat{\theta}_{ML}} = 0,$$
或
$$\left(\frac{\partial l(z \mid \theta)}{\partial \theta}\right)^{T}\Bigg|_{\hat{\theta}_{ML}} = 0.$$
上式为极大似然原理，其中 $\hat{\theta}_{ML}$ 称为**极大似然估计**，它实现了
$$p(z(k) \mid \hat{\theta}_{ML}) \xrightarrow{\max} p(z(k) \mid \theta_0).$$

6.2 极大似然参数估计

考虑如下模型
$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + D(Z^{-1})v(k), \tag{6-1}$$
其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$v(k)$ 是均值为零、方差为 σ_v^2，且服从正态分布的不相关随机噪声；$A(Z^{-1})$、$B(Z^{-1})$ 和 $D(Z^{-1})$ 为迟延算子多项式，分别记作：
$$\begin{cases} A(Z^{-1}) = 1 + a_1 Z^{-1} + a_2 Z^{-2} + \cdots + a_{n_a} Z^{-n_a}, \\ B(Z^{-1}) = b_1 Z^{-1} + b_2 Z^{-2} + \cdots + b_{n_b} Z^{-n_b}, \\ D(Z^{-1}) = 1 + d_1 Z^{-1} + d_2 Z^{-2} + \cdots + d_{n_d} Z^{-n_d}, \end{cases} \tag{6-2}$$
其中 n_a，n_b 和 n_d 为模型阶次. 又设 $e(k) = D(Z^{-1})v(k)$，其协方差阵记作 Σ_e. 对模型(6-1)式运用极大似然原理，可以获得如下结果：

1° 如果噪声 $e(k)$ 的协方差阵 Σ_e 已知，定义模型参数 $\theta = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b})^{T}$，则其极大似然估计为 $\hat{\theta}_{ML} = (H_L^{T}\Sigma_e^{-1}H_L)^{-1}H_L^{T}\Sigma_e^{-1}z_L$，其中数据矩阵 H_L 和输出向量 z_L 的定义与(3-7)式的相同. 这时的极大似然估计 $\hat{\theta}_{ML}$ 与马尔可夫估计是等价的.

2° 如果噪声 $e(k)$ 的协方差阵 $\Sigma_e = \sigma_v^2 I$，则模型参数 θ 的极大似然估计变为

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^{\mathrm{T}}\boldsymbol{z}_L,$$

这时的极大似然估计 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ 与最小二乘估计是等阶的 .

3° 如果噪声 $e(k)$ 的协方差阵 $\boldsymbol{\Sigma}_e$ 未知,定义模型参数 $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}$,则模型参数的极大似然估计问题可以归结为如下的非线性优化问题:

优化目标 $\quad V(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \dfrac{1}{L}\sum_{k=1}^{L} v^2(k)\Big|_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} \to \min;$

约束条件 $\quad v(k) = A(\mathrm{Z}^{-1})z(k) - B(\mathrm{Z}^{-1})u(k) - (D(\mathrm{Z}^{-1}) - 1)v(k).$

同时还可以获得噪声 $v(k)$ 方差的估计值

$$\tilde{\sigma}_v^2 = \min V(\boldsymbol{\theta}) = V(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}).$$

显然,优化目标函数 $V(\boldsymbol{\theta})$ 是参数 a_i, b_i, d_i 的函数,它关于 a_i, b_i 是线性的,而关于 d_i 是非线性的,因此,需要用解非线性优化问题算法进行迭代求解 .

6.3 极大似然递推算法

如果噪声 $e(k)$ 的协方差阵 $\boldsymbol{\Sigma}_e$ 未知,则(6-1)式模型参数 $\boldsymbol{\theta} = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, d_1, d_2, \cdots, d_{n_d})^{\mathrm{T}}$ 的极大似然估计递推算法为:

$$\begin{cases}
\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{K}(k)(z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1)),\\
\boldsymbol{K}(k) = \boldsymbol{P}(k-1)\boldsymbol{h}_f(k)(\boldsymbol{h}_f^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{h}_f(k) + 1)^{-1},\\
\boldsymbol{P}(k) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{h}_f^{\mathrm{T}}(k))\boldsymbol{P}(k-1),\\
\boldsymbol{h}(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a),\\
\qquad u(k-1), u(k-2), \cdots, u(k-n_b),\\
\qquad \hat{v}(k-1), \hat{v}(k-2), \cdots, \hat{v}(k-n_d))^{\mathrm{T}},\\
\boldsymbol{h}_f(k) = (-z_f(k-1), -z_f(k-2), \cdots, -z_f(k-n_a),\\
\qquad u_f(k-1), u_f(k-2), \cdots, u_f(k-n_b),\\
\qquad \hat{v}_f(k-1), \hat{v}_f(k-2), \cdots, \hat{v}_f(k-n_d))^{\mathrm{T}},\\
z_f(k) = z(k) - \hat{d}_1(k)z_f(k-1) - \hat{d}_2(k)z_f(k-2) -\\
\qquad \cdots - \hat{d}_{n_d}(k)z_f(k-n_d),\\
u_f(k) = u(k) - \hat{d}_1(k)u_f(k-1) - \hat{d}_2(k)u_f(k-2) -\\
\qquad \cdots - \hat{d}_{n_d}(k)u_f(k-n_d),\\
\hat{v}_f(k) = \hat{v}(k) - \hat{d}_1(k)\hat{v}_f(k-1) - \hat{d}_2(k)\hat{v}_f(k-2),\\
\qquad \cdots - \hat{d}_{n_d}(k)\hat{v}_f(k-n_d),\\
\hat{v}(k) = z(k) - \boldsymbol{h}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1),
\end{cases}$$

记作 RML(recursive maximum likelihood estimation algorithm).

6.4　极大似然估计的统计性质

当输入信号 $u(k)$ 满足 $2n(n = \max(n_a, n_b, n_d))$ 阶"持续激励"条件时,(6-1)式模型参数的**极大似然估计**具有如下良好的统计性质:

1° 一致性(相容性),即

$$\hat{\boldsymbol{\theta}}_{ML} \xrightarrow{L \to \infty} \boldsymbol{\theta}_0 \qquad (\text{a.s.}).$$

2° 渐近正态性,即

$$\sqrt{L}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{\text{law}} \beta \sim N(\boldsymbol{0}, \overline{\boldsymbol{M}}_{\theta_0}^{-1}),$$

其中 $\overline{\boldsymbol{M}}_{\theta_0}$ 为真实模型参数 $\boldsymbol{\theta}_0$ 条件下的平均费希尔信息矩阵.

3° 有效性,即 $\boldsymbol{E}\{(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)^{\mathrm{T}}\} = \frac{1}{L}\overline{\boldsymbol{M}}_{\theta_0}^{-1}$.

7　预报误差法

7.1　预报误差模型

考虑一般的模型类

$$z(k) = f(\boldsymbol{Z}^{k-1}, \boldsymbol{U}^k, k, \boldsymbol{\theta}) + \nu(k), \tag{7-1}$$

其中 $z(k) \in \mathbf{R}^m$ 为模型输出;$\boldsymbol{Z}^{k-1}$ 表示 $k-1$ 时刻以前的输出数据集合$\{z(k-1), z(k-2), \cdots\}$;$\boldsymbol{U}^k$ 表示 k 时刻以前的输入数据集合$\{\boldsymbol{u}(k), \boldsymbol{u}(k-1), \cdots\}$,$\boldsymbol{u}(k) \in \mathbf{R}^r$;$\boldsymbol{\theta}$ 为模型参数向量;$\{\nu(k)\}$ 为模型新息序列,在给定的数据集合 $\boldsymbol{Z}^{k-1}, \boldsymbol{U}^k$ 下,它具备条件均值等于零的性质,即 $E\{\nu(k) \mid \boldsymbol{Z}^{k-1}, \boldsymbol{U}^k\} = 0$. 这种模型称为预报误差模型(prediction error model,简称 PEM),它特别适合用于参数估计.

7.2　预报误差准则

在获得数据集合 Z^{k-1}, U^k 的条件下,对模型输出 $z(k)$ 的"最好"预报可取它的条件数学期望,即 $\hat{z}(k \mid \boldsymbol{\theta}) = E\{Z(k) \mid z^{k-1}, \boldsymbol{U}^k, \boldsymbol{\theta}\}$,它使

$$E\{\| z(k) - \hat{z}(k \mid \boldsymbol{\theta}) \|^2 \mid z^{k-1}, U^k, \boldsymbol{\theta}\} \to \min.$$

这种"最好"的输出预报应该是"最好"模型的输出. 对于特定的 $\boldsymbol{\theta}$ 值,(7-1)式模型的预报误差可写成

$$\tilde{z}(k, \boldsymbol{\theta}) = z(k) - \hat{z}(k \mid \boldsymbol{\theta}) = z(k) - f(\boldsymbol{Z}^{k-1}, \boldsymbol{U}^k, k, \boldsymbol{\theta}),$$

或写成拟线性回归的形式:

$$\tilde{z}(k,\boldsymbol{\theta}) = z(k) - \hat{z}(k \mid \boldsymbol{\theta}) = z(k) - \boldsymbol{\phi}^{\mathrm{T}}(k)\boldsymbol{\theta},$$

其中 $\boldsymbol{\phi}(k)$ 为由过去数据构成的数据向量．对应的预报误差序列$\{\tilde{z}(k,\boldsymbol{\theta})\}$的样本协方差为

$$D(\boldsymbol{\theta}) = \frac{1}{L}\sum_{k=1}^{L}\tilde{z}(k,\boldsymbol{\theta})\tilde{z}^{\mathrm{T}}(k,\boldsymbol{\theta}).$$

显然，一个好的模型应该具有比较小的预报误差，因此，可用 $D(\boldsymbol{\theta})$ 的某种正标量函数作为**预报误差准则**．通常选用的标量函数有

(1) $J_1(\boldsymbol{\theta}) = \mathrm{tr}(\boldsymbol{A}_L D(\boldsymbol{\theta}))$．如果 $v(k)$ 服从正态分布，均值为零，其协方差阵为 $\boldsymbol{\Sigma}_v(k)$，且预报误差准则函数 $J_1(\boldsymbol{\theta})$ 中的加权矩阵取为 $\boldsymbol{A}_L = L\boldsymbol{\Sigma}_v^{-1}(k)$，其中 L 为数据长度，则极小化 $J_1(\boldsymbol{\theta})$ 的结果与极大似然估计是等价的．

(2) $J_2(\boldsymbol{\theta}) = \ln \det D(\boldsymbol{\theta})$．如果 $v(k)$ 的协方差阵为 $\boldsymbol{\Sigma}_v(k)$ 未知，则极小化 $J_2(\boldsymbol{\theta})$ 的结果与极大似然估计是等价的．

预报误差准则 $J_1(\boldsymbol{\theta})$ 的另一种写法为

$$J(\boldsymbol{\theta}) = \tilde{z}^{\mathrm{T}}(k,\boldsymbol{\theta})\boldsymbol{A}_L\tilde{z}(k,\boldsymbol{\theta}).$$

为了去掉对样本的依赖性，引入如下函数作为预报误差准则：

$$J(\boldsymbol{\theta}) = \frac{1}{2}E\{\tilde{z}^{\mathrm{T}}(k,\boldsymbol{\theta})\boldsymbol{A}_L\tilde{z}(k,\boldsymbol{\theta})\}. \tag{7-2}$$

这是一种最常用的预报误差准则函数．

7.3　预报误差算法

极小化(7-2)式预报误差准则 $J(\boldsymbol{\theta})$，使预报误差变得尽可能的小．这种估计模型参数的方法就称为预报误差法，其递推形式称递推预报误差法，记作 **RPEM**(recursive prediction error method)．显然，预报误差算法与所用的模型输出预报器结构有关．对于给定的一种预报器结构，通过极小化 $J(\boldsymbol{\theta})$ 可以估计出该结构下的模型参数 $\boldsymbol{\theta}$．这个过程包括求出预报器的构成及对预报值关于 $\boldsymbol{\theta}$ 的求导，这是预报误差算法复杂性的根源．

7.3.1　预报误差法的一般结构

利用随机牛顿搜索方向，在准则函数(7-2)式意义下，(7-1)式模型参数估计算法可写成

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) - \rho(k)\boldsymbol{R}^{-1}(k)\left(\frac{\partial J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}\Bigg|_{\hat{\boldsymbol{\theta}}(k-1)},$$

其中 $\boldsymbol{R}(k)$ 是黑塞矩阵$\dfrac{\partial^2 J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}$ 的近似表达式．

若定义 $\boldsymbol{\Psi}(k,\boldsymbol{\theta}) = \left(\dfrac{\partial\hat{z}(k \mid \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}}$，则有

$$\begin{cases}\left(\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} = E\{\boldsymbol{\Psi}(k,\boldsymbol{\theta})\boldsymbol{\Lambda}_L\tilde{z}(k,\boldsymbol{\theta})\}, \\ \boldsymbol{R}(k) = \dfrac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = E\{\boldsymbol{\Psi}(k,\boldsymbol{\theta})\boldsymbol{\Lambda}_L\boldsymbol{\Psi}^{\mathrm{T}}(k,\boldsymbol{\theta})\}.\end{cases}$$

设加权矩阵

$$\boldsymbol{\Lambda}_L = \boldsymbol{\Sigma}_v^{-1}(k) = E\{(\tilde{z}(k,\boldsymbol{\theta})\tilde{z}^{\mathrm{T}}(k,\boldsymbol{\theta}))^{-1}\},$$

再利用随机逼近原理，对$\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$，$\boldsymbol{R}(k)$ 和 $\boldsymbol{\Sigma}_v^{-1}(k)$ 分别导出它们的递推计算形式，由此就可构成如下预报误差算法的一般结构形式：

$$\begin{cases}\tilde{z}(k,\hat{\boldsymbol{\theta}}(k-1)) = z(k) - \hat{z}(k\mid\hat{\boldsymbol{\theta}}(k-1)), \\ \hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \rho(k)\boldsymbol{R}^{-1}(k)\boldsymbol{\Psi}(k,\hat{\boldsymbol{\theta}}(k-1))\boldsymbol{\Sigma}_v^{-1}(k)\tilde{z}(k,\hat{\boldsymbol{\theta}}(k-1)), \\ \boldsymbol{R}(k) = \boldsymbol{R}(k-1) + \rho(k)(\boldsymbol{\Psi}(k,\hat{\boldsymbol{\theta}}(k-1))\boldsymbol{\Sigma}_v^{-1}(k)\boldsymbol{\Psi}^{\mathrm{T}}(k,\hat{\boldsymbol{\theta}}(k-1)) - \boldsymbol{R}(k-1)), \\ \hat{\boldsymbol{\Sigma}}_v(k) = \hat{\boldsymbol{\Sigma}}_v(k-1) + \rho(k)(\tilde{z}(k,\hat{\boldsymbol{\theta}}(k-1))\tilde{z}^{\mathrm{T}}(k,\hat{\boldsymbol{\theta}}(k-1)) - \hat{\boldsymbol{\Sigma}}_v(k-1)), \\ \boldsymbol{x}(k+1) = \boldsymbol{A}(\hat{\boldsymbol{\theta}}(k))\boldsymbol{x}(k) + \boldsymbol{B}(\hat{\boldsymbol{\theta}}(k))\bar{\boldsymbol{u}}(k), \\ \begin{pmatrix}\hat{z}(k\mid\hat{\boldsymbol{\theta}}(k)) \\ \mathrm{col}\boldsymbol{\Psi}^{\mathrm{T}}(k,\hat{\boldsymbol{\theta}}(k))\end{pmatrix} = \boldsymbol{C}(\hat{\boldsymbol{\theta}}(k))\boldsymbol{x}(k),\end{cases}$$

其中 $\rho(k)$ 为随机收敛因子；状态变量 $\boldsymbol{x}(k)$，系数矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 的定义见(2-9)式．上式表明，预报误差算法的关键在于求出预报器结构和预报值关于参数 $\boldsymbol{\theta}$ 的导数．预报器的结构与具体的模型形式有关．下面介绍三种特定模型的递推预报误差参数辨识算法．

7.3.2 ARMAX 模型的 RPEM

考虑 ARMAX 模型

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + D(Z^{-1})v(k),$$

其输出预报器为

$$\hat{z}(k\mid\boldsymbol{\theta}) = z(k) - \frac{A(Z^{-1})}{D(Z^{-1})}z(k) + \frac{B(Z^{-1})}{D(Z^{-1})}u(k).$$

置参数向量 $\boldsymbol{\theta} = (a_1,a_2,\cdots,a_{n_a},b_1,b_2,\cdots,b_{n_b},d_1,d_2,\cdots,d_{n_d})^{\mathrm{T}}$，取加权矩阵为单位阵，用 $\hat{z}(k)$ 表示输出预报值，$\tilde{z}(k)$ 表示预报误差，$\boldsymbol{\varphi}(k)$ 表示预报值关于参数 $\boldsymbol{\theta}$ 的一阶梯度，套用(7-3)式预报误差算法，可导出 ARMAX 模型的预报误差辨识算法为

$$\tilde{z}(k) = z(k) - \hat{z}(k),$$

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \rho(k)\boldsymbol{R}^{-1}(k)\boldsymbol{\varphi}(k)\tilde{z}(k),$$

$$\boldsymbol{R}(k) = \boldsymbol{R}(k-1) + \rho(k)(\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{\varphi}(k) - \boldsymbol{R}(k-1)),$$

$$D(Z^{-1})\mid \hat{z}_{\hat{\boldsymbol{\theta}}(k-1)}(k) = (D(Z^{-1}) - A(Z^{-1}))\mid z_{\hat{\boldsymbol{\theta}}(k-1)}(k) + B(Z^{-1})\mid u_{\hat{\boldsymbol{\theta}}(k-1)}(k),$$

$$D(Z^{-1}) \mid \varphi_{\hat{\theta}(k-1)}(k) = h(k),$$

$$h(k) = (-z(k-1), -z(k-2), \cdots, -z(k-n_a), u(k-1), u(k-2),$$
$$\cdots, u(k-n_b), \tilde{z}(k-1), \tilde{z}(k-2), \cdots, \tilde{z}(k-n_d))^{\mathrm{T}}.$$

7.3.3 SISO 模型的 RPEM

考虑一般的 SISO 模型

$$A(Z^{-1})z(k) = \frac{B(Z^{-1})}{F(Z^{-1})}u(k) + \frac{D(Z^{-1})}{C(Z^{-1})}v(k),$$

其输出预报器为

$$\hat{z}(k \mid \theta) = z(k) - \frac{A(Z^{-1})C(Z^{-1})}{D(Z^{-1})}z(k) + \frac{B(Z^{-1})C(Z^{-1})}{F(Z^{-1})D(Z^{-1})}u(k).$$

置参数向量

$$\theta = (a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}, c_1, c_2, \cdots, c_{n_c}, d_1, d_2, \cdots, d_{n_d}, f_1, f_2, \cdots, f_{n_f})^{\mathrm{T}},$$

取加权矩阵为单位阵,用 $\hat{z}(k)$ 表示模型输出预报值,$\tilde{z}(k)$ 表示预报误差,$\varphi(k)$ 表示预报值关于参数的一阶梯度,引入辅助变量

$$u^*(k) = \frac{B(Z^{-1})}{F(Z^{-1})}u(k),$$
$$v^*(k) = A(Z^{-1})z(k) - u^*(k),$$

则输出预报器可写成

$$\begin{cases} D(Z^{-1}) \mid \hat{z}_{\hat{\theta}(k-1)}(k) = D(Z^{-1}) \mid z_{\hat{\theta}(k-1)}(k) - C(Z^{-1}) \mid v^*_{\hat{\theta}(k-1)}(k), \\ v^*(k) = A(Z^{-1}) \mid z_{\hat{\theta}(k-1)}(k) - u^*(k), \\ F(Z^{-1}) \mid u^*_{\hat{\theta}(k-1)}(k) = B(Z^{-1}) \mid_{\hat{\theta}(k-1)} u(k). \end{cases}$$

输出预报值关于参数的一阶梯度向量 $\varphi(k)$ 由下列元素组成:

$$\begin{cases} \dfrac{\partial \hat{z}(k)}{\partial a_i} = -\dfrac{C(Z^{-1})}{D(Z^{-1})}z(k-i) & (i = 1,2,\cdots,n_a), \\ \dfrac{\partial \hat{z}(k)}{\partial b_i} = \dfrac{C(Z^{-1})}{D(Z^{-1})F(Z^{-1})}u(k-i) & (i = 1,2,\cdots,n_b), \\ \dfrac{\partial \hat{z}(k)}{\partial f_i} = -\dfrac{C(Z^{-1})}{D(Z^{-1})F(Z^{-1})}u^*(k-i) & (i = 1,2,\cdots,n_f), \\ \dfrac{\partial \hat{z}(k)}{\partial c_i} = -\dfrac{1}{D(Z^{-1})}v^*(k-i) & (i = 1,2,\cdots,n_c), \\ \dfrac{\partial \hat{z}(k)}{\partial d_i} = \dfrac{1}{D(Z^{-1})}\tilde{z}(k-i) & (i = 1,2,\cdots,n_d). \end{cases}$$

由此可构成一般模型的预报误差辨识算法:

$$\begin{cases}\tilde{z}(k)=z(k)-\hat{z}(k),\\ \hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\rho(k)\boldsymbol{R}^{-1}(k)\boldsymbol{\varphi}(k)\tilde{z}(k),\\ \boldsymbol{R}(k)=\boldsymbol{R}(k-1)+\rho(k)(\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{\varphi}(k)-\boldsymbol{R}(k-1)),\end{cases}$$

其中 $\rho(k)$ 为收敛因子．

这种算法相当一般化．当迟延算子多项式 $A(Z^{-1}),B(Z^{-1}),C(Z^{-1}),D(Z^{-1})$ 和 $F(Z^{-1})$ 中某几个多项式为1时,上述算法可直接演化成多种常用算法．比如,当 $C(Z^{-1})=F(Z^{-1})=1$ 时就演化成极大似然递推算法．

7.3.4 线性新息模型的 RPEM

考虑如下线性状态空间模型:

$$\begin{cases}\boldsymbol{x}(k+1)=\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}(k)+\boldsymbol{b}(\boldsymbol{\theta})u(k)+v_1(k),\\ z(k)=\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{x}(k)+v_2(k),\end{cases}$$

其中 $u(k),z(k)$ 分别为模型输入和输出变量;$\boldsymbol{x}(k)$ 为模型状态变量;$\boldsymbol{\theta}$ 为模型参数;$v_1(k)$、$v_2(k)$ 为不相关的零均值白噪声．

根据卡尔曼(R.E.Kalman)滤波器原理,模型输出预报值 $\hat{z}(k)$ 可写成

$$\begin{cases}\hat{\boldsymbol{x}}(k+1)=\boldsymbol{A}(\hat{\boldsymbol{\theta}}(k-1))\hat{\boldsymbol{x}}(k)+\boldsymbol{b}(\hat{\boldsymbol{\theta}}(k-1))u(k)+\\ \qquad\qquad \boldsymbol{K}(k)(z(k)-\boldsymbol{c}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}(k-1))\hat{\boldsymbol{x}}(k)),\\ \hat{z}(k)=\boldsymbol{c}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}(k-1))\hat{\boldsymbol{x}}(k),\end{cases}$$

其中 $\boldsymbol{K}(k)$ 为卡尔曼增益;预报误差 $\tilde{z}(k)=z(k)-\hat{z}(k)$;预报值关于参数 $\boldsymbol{\theta}$ 的一阶梯度 $\boldsymbol{\varphi}(k)$ 的第 i 个元素为

$$\varphi_i(k)=\frac{\partial\boldsymbol{c}(\hat{\boldsymbol{\theta}}(k-1))}{\partial\theta_i}\hat{\boldsymbol{x}}(k)+\boldsymbol{c}(\hat{\boldsymbol{\theta}}(k-1))\frac{\partial\hat{\boldsymbol{x}}(k)}{\partial\theta_i},$$

其中

$$\frac{\partial\hat{\boldsymbol{x}}(k)}{\partial\theta_i}=\frac{\partial\boldsymbol{A}(\hat{\boldsymbol{\theta}}(k-1))}{\partial\theta_i}\hat{\boldsymbol{x}}(k-1)+\boldsymbol{A}(\hat{\boldsymbol{\theta}}(k-1))\frac{\partial\hat{\boldsymbol{x}}(k-1)}{\partial\theta_i}+$$

$$\frac{\partial\boldsymbol{B}(\hat{\boldsymbol{\theta}}(k-1))}{\partial\theta_i}u(k)+\frac{\partial\boldsymbol{K}(k)}{\partial\theta_i}\tilde{z}(k-1)-\boldsymbol{K}(k-1)\varphi_i(k-1),$$

那么状态空间模型的预报误差辨识算法可写成

$$\begin{cases}\tilde{z}(k)=z(k)-\hat{z}(k),\\ \hat{\boldsymbol{\theta}}(k)=\hat{\boldsymbol{\theta}}(k-1)+\rho(k)\boldsymbol{R}^{-1}(k)\boldsymbol{\varphi}(k)\tilde{z}(k),\\ \boldsymbol{R}(k)=\boldsymbol{R}(k-1)+\rho(k)(\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{\varphi}(k)-\boldsymbol{R}(k-1)),\end{cases}$$

其中 $\rho(k)$ 为收敛因子．

第3章至第6章讨论了三类辨识算法,它们是最小二乘类方法、梯度校正法和极大似然法．其中最小二乘法是最常用的辨识方法,其收敛速度快,鲁棒性强．可

能的话,应当尽量使用最小二乘类算法．第7章还讨论了预报误差法的辨识框架．这种方法的原理非常一般,可用于更广泛的辨识问题．

8　模型结构辨识

各种模型参数辨识方法一般需要假定模型的结构已知,但实际上在多数情况下这是不现实的．当没有模型结构的先验知识时,需要利用系统的输入、输出数据来确定模型的结构．这就是所谓的**模型结构辨识问题**．对单输入单输出(SISO)系统来说,模型结构辨识也就是模型阶次辨识．下面介绍各种模型结构辨识方法．

8.1　残差方差检验法

考虑如下模型:

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + v(k). \tag{8-1}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量;$v(k)$ 是均值为零、方差为 σ_v^2 不相关随机噪声;$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式,记为

$$\begin{cases} A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_nZ^{-n}, \\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_nZ^{-n}, \end{cases}$$

其中 n 为模型阶次．

模型(8-1)式的最小二乘格式可写成

$$z(k) = \boldsymbol{h}_n^{\mathrm{T}}(k)\boldsymbol{\theta}_n + v(k), \tag{8-2}$$

其中数据向量和参数向量分别定义为

$$\begin{cases} \boldsymbol{h}_n(k) = (-z(k-1),u(k-1),-z(k-2),u(k-2),\cdots,-z(k-n),u(k-n))^{\mathrm{T}}, \\ \boldsymbol{\theta}_n = (a_1,b_1,a_2,b_2,\cdots,a_n,b_n)^{\mathrm{T}}, \end{cases}$$

对(8-2)式运用最小二乘原理,可获得模型参数 $\boldsymbol{\theta}_n$ 的最小二乘估计为

$$\hat{\boldsymbol{\theta}}_n = (\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{H}_n)^{-1}\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{z}_n,$$

数据矩阵和输出向量分别定义为

$$\boldsymbol{H}_n = \begin{bmatrix} \boldsymbol{h}_n^{\mathrm{T}}(1) \\ \boldsymbol{h}_n^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{h}_n^{\mathrm{T}}(L) \end{bmatrix}, \quad \boldsymbol{z}_n = \begin{bmatrix} z(1) \\ z(2) \\ \vdots \\ z(L) \end{bmatrix},$$

其中 L 为数据长度.

当模型阶次为 n 时,输出残差向量可写成

$$\boldsymbol{\varepsilon}_n = \boldsymbol{z}_n - \boldsymbol{H}_n\hat{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\chi}}_n + \boldsymbol{v}_n,$$

其中

$$\begin{cases} \tilde{\chi}_n = H_{n0}\theta_{n0} - H_n\hat{\theta}_n, \\ v_n = (v(1), v(2), \cdots, v(L))^{\mathrm{T}}. \end{cases}$$

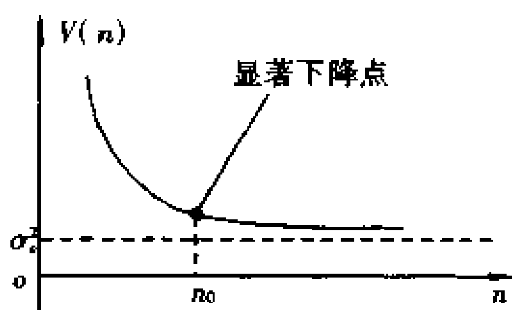


图 8-1

且残差的方差 $V(n)$ 具有如下性质：

$$P\lim_{L\to\infty} V(n) = P\lim_{L\to\infty}\left(\frac{1}{L}\varepsilon_n^{\mathrm{T}}\varepsilon_n\right) = P\lim_{L\to\infty}\left(\frac{1}{L}\tilde{\chi}_n^{\mathrm{T}}\tilde{\chi}_n\right) + P\lim_{L\to\infty}\left(\frac{1}{L}v_b^{\mathrm{T}}v_n\right),$$

它具有如图8-1所示的变化特性．这样，通过观察残差方差 $V(n)$ 的变化情况，可确定模型的阶次，具体步骤如下：

(1) 阶次 n 逐一增加，当 n 增加至 $\hat{n}$ 时，若 $V(\hat{n})$ 呈现显著下降趋势，则确认模型阶次为 $\hat{n}$.

(2) 引进统计量

$$t = \frac{V(n_1) - V(n_2)}{V(n_2)}\frac{L - 2n_2}{2(n_2 - n_1)} \sim F(2(n_2 - n_1), L - 2n_2),$$

t 服从 F 分布，自由度为 $2(n_2 - n_1)$ 和 $L - 2n_2$.

(3) 设零假设为

$$H_0: n_2 > n_1 \geqslant n_0,$$

取风险水平 $\alpha = 5\%$，则对应的阈值 $t_\alpha = F(2(n_2 - n_1), L - 2n_2)$.

(4) 按下式判断模型阶次：

当 $t(n_2, n_1) > t_\alpha$ 时， 拒绝 $H_0: n_1 \geqslant n_0$；

当 $t(n_2, n_1) \leqslant t_\alpha$ 时， 接受 $H_0: n_2 \geqslant n_0$.

这时模型阶次的估计值应为 $\hat{n} = n_2$.

(5) 如果模型(8-1)式噪声是有色噪声，引入噪声模型描述后，定阶的方法与上述类似．

8.2 AIC定阶法

8.2.1 AIC准则

考虑如下线性模型：

$$z(k) = h_1(k)\theta_1 + h_2(k)\theta_2 + \cdots + h_N(k)\theta_N + e(k).$$

其中 $z(k)$ 为模型输出；$h_i(k), i=1,2,\cdots,N$，为 N 个独立的模型输入变量；$\theta_i, i=1,2,\cdots,N$，为 N 个独立的模型参数；$e(k)$ 为模型噪声．为了确定这类模型的独立参数个数(相当于模型阶次)，赤池(Akaike)引进如下准则：

$$\mathrm{AIC}(N) = -2\ln L(\hat{\boldsymbol{\theta}}) + 2N,$$

其中 N 是模型的参数个数

$$\begin{cases} N = \dim\boldsymbol{\theta}, \\ \boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)^{\mathrm{T}}, \end{cases}$$

$L(\hat{\boldsymbol{\theta}})$ 是 $\hat{\boldsymbol{\theta}}$ 条件下的似然函数．这个准则通常称为 **AIC**(Akaike information criterion) **准则**．当 AIC(N) 达到最小时，对应的 N 可为模型相对合理的阶次估计值．

8.2.2　定性解释

当模型阶次 N 低于真实阶次 N_0 时，AIC 准则中的似然函数 $L(\hat{\boldsymbol{\theta}})$ 将随着 N 的增加而增大，这时 AIC 准则的 AIC(N) 呈现下降趋势．当模型阶次 N 超过真实阶次 N_0 时，似然函数 $L(\hat{\boldsymbol{\theta}})$ 增长的趋势将大大放慢，这是因为模型已经接近真实系统，似然函数 $L(\hat{\boldsymbol{\theta}})$ 不再会有大的变化．这时 N 的增加速度会超过似然函数 $L(\hat{\boldsymbol{\theta}})$ 的增长速度，从而使 AIC 准则的 AIC(N) 呈现上升趋势．为此，AIC 准则一定存在极小值，如图 8-2 所示．

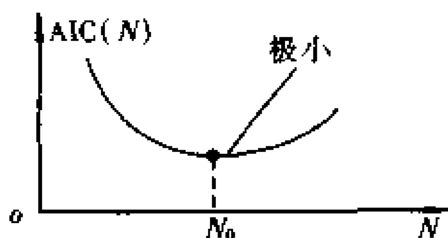


图 8-2

以使 AIC 准则达到极小的 N 作为模型阶次的估计值，或者说以此确定模型的独立参数个数，这种模型阶次辨识方法称作 **AIC 定阶法**．

8.2.3　AIC 定阶法

1. 白噪声情况

考虑如下模型：

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + v(k). \tag{8-3}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$v(k)$ 是均值为零、方差为 σ_v^2、服从正态分布的不相关随机噪声；$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式，分别记作

$$\begin{cases} A(Z^{-1}) = 1 + a_1 Z^{-1} + a_2 Z^{-2} + \cdots + a_{n_a} Z^{-n_a}, \\ B(Z^{-1}) = b_1 Z^{-1} + b_2 Z^{-2} + \cdots + b_{n_b} Z^{-n_b}, \end{cases}$$

其中 n_a 和 n_b 为模型阶次．

依据所给条件，可导出模型(8-3)式的 AIC 准则为

$$\mathrm{AIC}(n_a, n_b) = L\ln\widehat{\sigma_v^2} + 2(n_a + n_b),$$

其中 $\widehat{\sigma_v^2}$ 为噪声方差估计值．选择使 $\mathrm{AIC}(n_a, n_b)$ 达到最小的 n_a, n_b 作为模型的阶次．

2. 有色噪声情况

考虑如下模型：

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + D(Z^{-1})v(k). \tag{8-4}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$v(k)$ 是均值为零、方差为 σ_v^2、服从正态分布的不相关随机噪声；$A(Z^{-1})$，$B(Z^{-1})$ 和 $D(Z^{-1})$ 为迟延算子多项式，记作

$$\begin{cases} A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_{n_a}Z^{-n_a}, \\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_{n_b}Z^{-n_b}, \\ D(Z^{-1}) = 1 + d_1Z^{-1} + d_2Z^{-2} + \cdots + d_{n_d}Z^{-n_d}, \end{cases}$$

其中 n_a，n_b 和 n_d 为模型阶次．

依据所给条件，可导出(8-4)式 AIC 准则为

$$\mathrm{AIC}(n_a, n_b, n_d) = L\ln\widehat{\sigma_v^2} + 2(n_a + n_b + n_d),$$

其中 $\widehat{\sigma_v^2}$ 为噪声方差估计值．选择使 $\mathrm{AIC}(n_a, n_b, n_d)$ 达到最小的 n_a，n_b 和 n_b 作为模型的阶次．

8.3　最终预报误差法

8.3.1　概念

AIC 定阶法需要确定似然函数，这就要求模型噪声的概率分布必须为已知．如果模型噪声的概率分布无法知道，则需要用最终预报误差准则来估计模型的阶次．应该说，“最好”的模型能给出“最好”的输出预报值，也就说最终预报误差准则达到最小的模型是“最好”的模型．根据这一道理，当模型阶次逐一增加，最终预报误差准则达到最小时，对应的阶次可当做模型阶次估计值．这种模型阶次辨识方法称作 FPE 定阶法．

8.3.2　FPE 定阶法

1. 白噪声情况

考虑如下模型：

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + v(k). \tag{8-5}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量；$v(k)$ 是均值为零、方差为 σ_v^2 的不相关随机噪声；$A(Z^{-1})$ 和 $B(Z^{-1})$ 为迟延算子多项式，记作

$$\begin{cases}A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_{n_a}Z^{-n_a},\\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_{n_b}Z^{-n_b},\end{cases}$$

其中 n_a 和 n_b 为模型阶次.

依据所给条件,可导出模型(8-5)式的 FPE 准则为

$$\mathrm{FPE}(n_a, n_b) = \frac{L + (n_a + n_b)}{L - (n_a + n_b)}\hat{\sigma}_v^2,$$

其中 $\hat{\sigma}_v^2$ 为噪声方差估计值. 选择使 $\mathrm{FPE}(n_a, n_b)$ 达到最小的 n_a, n_b 作为模型的阶次.

2. 有色噪声情况

考虑如下模型:

$$A(Z^{-1})z(k) = B(Z^{-1})u(k) + D(Z^{-1})v(k). \tag{8-6}$$

其中 $u(k)$ 和 $z(k)$ 分别为模型输入和输出变量;$v(k)$ 是均值为零、方差为 σ_v^2 的不相关随机噪声;$A(Z^{-1})$, $B(Z^{-1})$ 和 $D(Z^{-1})$ 为迟延算子多项式,记作

$$\begin{cases}A(Z^{-1}) = 1 + a_1Z^{-1} + a_2Z^{-2} + \cdots + a_{n_a}Z^{-n_a},\\ B(Z^{-1}) = b_1Z^{-1} + b_2Z^{-2} + \cdots + b_{n_b}Z^{-n_b},\\ D(Z^{-1}) = 1 + d_1Z^{-1} + d_2Z^{-2} + \cdots + d_{n_d}Z^{-n_d},\end{cases}$$

其中 n_a, n_b 和 n_d 为模型阶次.

依据所给条件,可导出模型(8-6)式的 FPE 准则为

$$\mathrm{FPE}(n_a, n_b, n_d) = \frac{1 + \alpha - 2\beta + \gamma - 2\delta + \lambda}{1 - \alpha + 2\beta - \gamma + 2\delta - \lambda}\hat{\sigma}_v^2,$$

其中 $\hat{\sigma}_v^2$ 为噪声方差估计值;且

$$\begin{cases}\alpha = \sum_{i=1}^{n_a}\sum_{j=1}^{n_a}\frac{\partial^2 J(\boldsymbol{\theta})}{\partial a_i \partial a_j}R_z(j-i), & \beta = \sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\frac{\partial^2 J(\boldsymbol{\theta})}{\partial a_i \partial b_j}R_{uz}(j-i),\\ \gamma = \sum_{i=1}^{n_b}\sum_{j=1}^{n_b}\frac{\partial^2 J(\boldsymbol{\theta})}{\partial b_i \partial b_j}R_u(j-i), & \delta = \sum_{i=1}^{n_a}\sum_{j=1}^{n_d}\frac{\partial^2 J(\boldsymbol{\theta})}{\partial a_i \partial d_j}R_{z\hat{v}}(j-i),\\ \lambda = \sum_{i=1}^{n_d}\sum_{j=1}^{n_d}\frac{\partial^2 J(\boldsymbol{\theta})}{\partial d_i \partial d_j}R_{\hat{v}}(j-i),\end{cases}$$

其中 $\hat{v}(k)$ 为噪声 $v(k)$ 估计值;$J(\boldsymbol{\theta})$ 为辨识参数的准则函数,定义 $J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{L}v^2(k)$;$L$ 为数据长度;$R_{**}(j-i)$ 为对应的相关函数. 选择使 $\mathrm{FPE}(n_a, n_b, n_d)$ 达到最小的 n_a, n_b 和 n_d 当做模型阶次估计值.

9 系统辨识的试验设计

辨识就是根据含有噪声的输入、输出数据,从一类模型中确定与系统特性等价

的模型的过程．在实际应用中，除了合理选择模型类和辨识方法外，还有许多准备工作需要做．比如，掌握辨识对象的先验知识，明确辨识的目的，确定辨识实验方案和辨识模型的验证等．

9.1 可辨识性

9.1.1 可辨识性概念

如果根据系统的输入输出或状态变量的测量可以唯一地确定状态空间模型的系数矩阵或差分方程的迟延多项式参数，则系统是可辨识的．

不可控或不可观的系统是不可辨识的．因为这时可控性矩阵或可观性矩阵不满秩，使系统的外部描述仅依存于那些可控可观的状态，所以待辨识的系数矩阵中那些属于不可控或不可观状态的未知参数是无法利用系统的外部可测信号确定的．

9.1.2 可辨识性条件

开环系统的可辨识性除了要求系统模型是完全可控和可观的以外，辨识所用的输入信号必须能"持续激励"系统的所有模态．

闭环系统的可辨识性可归纳如下：

(1) 如果反馈通道不存在扰动信号，但反馈通道的模型阶次不低于前向通道的模型阶次，则系统是可辨识的．反馈通道或前向通道含有纯迟延将有利于闭环可辨识性条件的满足．

(2) 如果反馈通道拥有 $l \geqslant 1 + \frac{r}{m}$（$r$ 为输入个数，m 为输出个数）种不同参数的模型结构，且它们相互切换工作，则闭环系统是可辨识的．

(3) 如果反馈通道存在扰动信号，则系统是可辨识的．

(4) 如果反馈通道模型是非线性或时变的，则系统是可辨识的．

9.2 实验设计

辨识实验设计包括选择测量哪些信号，什么时候测量它们，以及控制哪些信号和怎么控制它们，使所生成的数据是充分提供信息的．

当面临着要辨识的实际系统时，首先遇到的问题是选择哪些信号作为输出？哪些信号作为输入？在实验期间用什么信号去激励系统？其次是选择数据采样间隔和输入激励信号．关于输入信号的选择有两个问题需要考虑：

(1) 输入信号的二阶统计矩性质，

(2) 输入信号的形状，常用的输入信号有正弦信号的叠加、白化后的噪声或伪随机信号等．

9.2.1　输入信号的选择

在辨识实验期间，输入信号必须能充分激励系统的所有模态．从谱分析角度看，这意味着输入信号的频谱必须足以覆盖系统的频谱．这就是所谓的持续激励问题．更进一步的要求是选择一种输入信号使给定问题的辨识模型精度最高．这种具有"优良性"的输入信号称为最优输入信号，它使费希尔信息矩阵的逆的某种标量函数达到最小．这个标量函数是用来评价模型精度的度量函数，记作

$$J = \phi(M^{-1}),$$

其中 M 为费希尔信息矩阵，

$$M = E\left\{\left[\frac{\partial \ln p(z \mid \theta)}{\partial \theta}\right]^{T}\left[\frac{\partial \ln p(z \mid \theta)}{\partial \theta}\right]\right\}.$$

上两式中，z 表示系统输出数据 $\{z(k), k = 1,2,\cdots,L\}$ 的集合；$p(z \mid \theta)$ 表示在模型参数 θ 条件下 z 的条件概率密度；ϕ 为某种标量函数，典型的有

(1) A 优化准则　$J = \mathrm{tr}[M^{-1}]$，

(2) D 优化准则　$J = \det M^{-1}$.

如果系统的输出数据是独立同分布的高斯(G. F. Gauss)随机序列，那么使 D 优化准则达到最小的输入信号是具有脉冲式自相关函数的信号，白噪声或 M 序列信号可以满足这一要求．

就工程意义上说，输入信号的选择还要考虑如下一些要求：

(1) 输入信号的功率或幅度不宜过大，以免系统工况进入非线性区；

(2) 输入信号对系统的"净扰动"要小；

(3) 工程上容易实现．

9.2.2　采样间隔的选择

采样间隔的选择至少需要考虑如下两个因素：

(1) 满足采样定理，即采样速度不能低于数据信号截止频率的 2 倍；

(2) 与模型应用的采样间隔时间尽可能保持一致．

如果系统的主时间常数记作 τ，则采样间隔的最优选择应该位于 τ 的附近．特别应该指出的是，使用太大的采样间隔比使用太小的采样间隔的辨识效果要坏得多．例如采用 10τ 的采样间隔，所产生的参数估计值偏差的方差将是采样间隔等于 τ 的 10^5 倍，而采用 0.1τ 的采样间隔所产生的参数估计值偏差的方差不会大于采样间隔等于 τ 的 10 倍．

9.2.3　数据的预处理

辨识实验收集到的观测数据可能含有高频扰动，也可能含有偶然的爆发、异常值、漂移和调零偏差或周期性低频扰动．高频扰动通常是因为采样间隔和预处理滤波器选择得不够理想引起的．爆发和异常值可以通过选择好的准则函数，或用故障检测算法去掉坏数据来克服．漂移、调零偏差或周期性低频扰动除了可采用零均值化和差分数据法消除外，还可以采用调整噪声模型结构的方法来消除．

9.3 模型结构的选择

模型结构的选择对成功的辨识十分重要．选择一种正确的模型结构至少包括如下三项内容：

(1) 选择模型类型，如选择线性模型或非线性模型；

(2) 选择模型阶次，包括准备将哪些变量包含在模型的描述中；

(3) 模型集的参数化，如果 $J(\theta)$ 为辨识的准则函数，则要求 $J(\theta)$ 关于参数 θ 是可微的．

一种合理的模型选择需要同时兼顾模型的灵活性和悭吝性．灵活性表现为模型需要包含更多的参数或把参数放在"关键位置"上，使之具有描述不同模型的能力，且使模型的偏差更小．悭吝性表现为模型不要包含不必要的参数，否则模型偏差的方差会随着参数个数的增加而增大．模型结构的选择还会影响辨识算法的复杂性，这是因为计算预报值和预报值关于参数 θ 的梯度时，所用的计算量与模型结构有关．此外，模型的预期用途和准则函数的性质也将影响模型结构的选择．有时同一个系统，可能需要选择多个模型结构，用以描述不同操作点或不同时间尺度下系统状况．选择模型结构的一个基本原则是："先试简单的"．如线性情形下，选择ARX 模型；非线性情形下，尽可能变换成线性表达式．

9.4 准则函数的选择

在系统辨识中，通常用某种准则函数来衡量模型的优劣，比如用某段时间内模型输出与系统输出之差的平方和，或样本均方 d 步超前预报误差，或极大似然准则等作为准则函数．准则函数的选择不仅影响辨识的精度，且与辨识算法的鲁棒性有关．如果采用极大似然准则作为辨识的准则函数，则可保证参数估计值偏差的协方差阵达到克拉默-拉奥不等式下界．如果系统噪声服从高斯正态分布，则以模型输出与系统输出之差的平方和作为准则函数，在无坏数据情况下可获得满意的辨识精度．要求辨识算法具有鲁棒性意味着算法对一些反常数据，或称坏数据要有较强的适应能力．合理地选择准则函数，可以降低辨识算法对坏数据的灵敏度，达到提高辨识算法鲁棒性的目的．

9.5 模型检验

模型检验就是评价辨识所获得的模型是否"足够好"，一般包括以下几个方面：

(1) 模型的输出是否与观测数据充分一致；

(2) 对建模的预期目的，模型是否足够好；

(3) 模型是否描述真实系统．

要回答上述三个问题，必须在模型中使用与实际一样多的信息，包括验前知识、实验数据、使用者的经验等，因此，要准确回答上述问题不是容易的事．一种简

单综合的方法是，检验模型残差的统计性质，或残差与过去的输入数据是否独立，这是因为在计算模型预报值$\hat{z}(k \mid \boldsymbol{\theta})$时，假设模型驱动噪声是独立随机噪声．下面几种方法可用于检验模型残差的独立性或称白色性，以及残差与过去输入数据是否独立．

（1）设$\{\varepsilon(k), k = 1,2,\cdots,L\}$代表模型残差序列，$L$为数据长度，令$\varepsilon(k)$的相关系数为

$$\rho_\varepsilon(l) = \frac{R_\varepsilon(k)}{R_\varepsilon(0)},$$

其中$R_\varepsilon(\cdot)$代表$\varepsilon(k)$的相关函数.

$$R(k) = \frac{l}{L}\sum_{k=1}^{L-l}\varepsilon(k)\varepsilon(k+1),$$

当L充分大时，若$\{\varepsilon(k)\}$确实是白噪声序列，则

$$t = L\sum_{l=1}^{m}\rho_\varepsilon^2(l) \sim \chi^2(m),$$

因此，可用$t \leqslant \chi_\alpha^2(m)$来检验模型残差$\{\varepsilon(k)\}$的独立性和白色性，其中$\alpha$为风险水平．若$L$比较大，则$m$取20～30即可满足工程上的要求．

（2）若满足下列条件：

$$\begin{cases} E\{\varepsilon(k)\} = 0, \\ |\rho_\varepsilon(k)| \leqslant \dfrac{1.98}{\sqrt{L}} & (l = 1,2,\cdots,20), \\ \text{或 } L\sum_{l=1}^{m}\rho_\varepsilon^2(l) \leqslant m + 1.65\sqrt{2m} & (m = 20 \sim 30), \end{cases}$$

则$\{\varepsilon(k), k = 1,2,\cdots,L\}$为白噪声序列．

（3）若模型残差$\varepsilon(k)$和输入信号$u(k)$的互相关函数满足

$$|R_{\varepsilon u}(l)| \leqslant \sqrt{\frac{\sum_{s=-L}^{L}R_\varepsilon(s)R_u(s)}{L}N_\alpha},$$

其中，$R_\varepsilon(s) = \overline{E}\{\varepsilon(k)\varepsilon(k+s)\}$，　$R_u(s) = \overline{E}\{u(k)u(k+s)\}$，

N_α表示正态分布$N(0,1)$当风险水平为α时的阈值，L为数据长度，则模型残差$\varepsilon(k)$和输入信号$u(k)$相互独立．

最后要强调，模型检验的客观标准应该是模型的实际应用效果．

参 考 文 献

1 ByoungSeon C. ARMA model identification. New York: Springer-Verlag, 1992.

2 Hellendoorn H, Driankov D. Fuzzy model identification: selected approaches. Berlin: Springer, 1997.

3 Isermann R. Digital control systems, Volume Ⅱ, Stochastic control, multivariable control,

adaptive control and applications. New York: Springer-Verlag, 1991.
4　Ljung L, Soderstrom T. Theory and practice of recursive identification. Cambridge: MIT Press, 1983.
5　(澳)Goodwin G C, Sin K S. 自适应滤波、预报与控制. 张永光等译. 北京:科学出版社, 1992.
6　(瑞典)Ljung L. 系统辨识—使用者的理论. 袁震东等译. 上海:华东师范大学出版社, 1990.
7　(日)相良节夫等著. 系统辨识. 萧德云等译. 北京:化学工业出版社, 1988.
8　方崇智, 萧德云著. 过程辨识. 北京:清华大学出版社, 1988.
9　刘宏才主编. 系统辨识与参数估计. 北京:冶金工业出版社, 1996.
10　张化光著. 复杂系统的模糊辨识与模糊自适应控制. 沈阳:东北大学出版社, 1993.

·经济数学卷·

第16篇

大系统理论

编　者　李人厚
审校者　秦寿康

目　录

引　言

自20世纪70年代开始,大系统理论逐渐形成为一个专门的学科.它综合了现代控制理论、控制论、图论、运筹学和决策论等方面的成果,是系统工程理论基础之一.大系统一般是指规模庞大(模型的维数很高)、结构复杂(多层次、互关联)、目标众多(目标间有冲突)、时标各异(同一系统内有多个时标)、地理位置分散,并常常具有随机性和不确定性的复杂系统.它不仅把复杂的工业系统作为研究对象,而且已扩展应用到社会、政治、经济和生态环境等系统中.

大系统理论涉及大系统模型简化、大系统结构、大系统稳定性以及大系统的递阶和分散控制等理论.国内外已把大系统理论成功地应用于电力系统、城市交通网、数字通信网络、计算机集成制造系统(CIMS)、生态系统、水资源系统和社会经济系统等各方面.

本篇将结合工程应用背景,从大系统结构入手,简要地介绍大系统理论的主要概念和算法。

1　大系统结构

大系统结构取决于组成大系统的子系统集合和各子系统之间的关联.大系统的结构决定了大系统的功能.不同的结构就产生不同的总体功能.由于大系统的受控对象分散,变量数目多,关联复杂,不宜采用集中式结构.在各种工程和非工程的大系统中,存在着两种基本结构:**递阶结构**(层次结构)和**分散结构**.

(1)递阶结构　在递阶结构中,整个大系统分成独立平行处理的许多子系统,且用一个协调器来协调各子系统之间的关联,通过上级(协调器)和下级间反复的信息交换实现协调过程.原则上协调器可以拥有局部控制器所有的全部信息.所以递阶结构具有"**经典信息模式**".

(2)分散结构　在分散结构中,各子系统独立工作,整个系统不存在协调器.各子系统只拥有局部信息,它只能通过各子系统之间的信息交换来调整总体目标,故分散结构具有"**非经典信息模式**".

按研究问题的目的,系统的分解和分级有不同的方式.从控制的角度出发,递阶结构又可分为多重递阶结构、多层递阶结构和多级递阶结构等.

1.1　多重递阶结构

多重递阶结构如图1-1所示.

所谓**多重递阶结构**是指用一组模型,从不同的抽象程度来对系统进行描述的

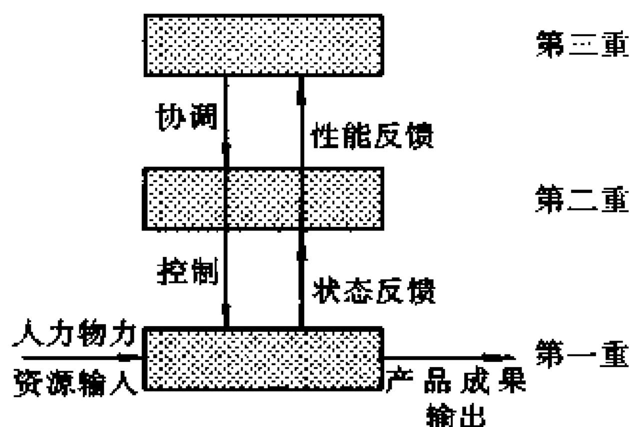


图 1-1

结构，这就形成了不同的层次，每一重都有相应的描述系统行为的变量、需要服从的规律，等等．例如对于一个复杂的自动化工业过程，可以按以下三重来研究(见图1-1)：

第一重，把系统看成是按照一定的物理规律变化的物理对象．

第二重，从信息处理和控制的角度出发，把过程看成是一个受控的系统．

第三重，从经济学的角度出发，把系统看成一个经济实体，它涉及到评价系统的效益和利润．

1.2　多层递阶结构

多层递阶结构是按系统决策的复杂性来分级的．例如，对于具有不确定因素的复杂系统，控制功能可以按四个层次来实现，如图 1-2 所示．

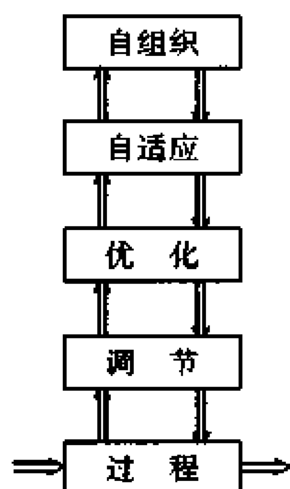


图 1-2

(1) 直接控制层或调节层　　它的任务是面对扰动的影响，力图把过程的有关变量维持在预先给定的设定值上．

(2) 优化层或监控层　它的任务是按照一定的最优性指标来规定直接控制层各控制器的设定值.解决这个问题通常都要用到对象的模型,而其中的某些参数是不确定的,需要上一级来设定.

(3) 学习层或自适应层　它的任务是根据对实际系统的观测来辨识优化层中所使用的模型的参数,使得模型尽量和变化的实际过程保持一致.

(4) 自组织层　它的任务是按照系统控制的总目标来选择下层所采用的模型结构、控制策略,等等.如果总目标有了变化,那么,它可以自动改变优化层中所用的性能指标,或者当参数辨识不能令人满意时它可以修改适应层的学习策略.

1.3　多级递阶结构

多级递阶结构系统由若干个明显可分的相互关联的子系统组成.所有的决策单元按一定的支配关系递阶排列.同一级的各个单元要受到上一级的干预,同时又对下一级的决策单元施加影响.虽然各自都有一定的自由度,但一般同级的决策单元可能会有互相冲突的决策目标,这就需要上级的协调.

按系统的复杂程度和控制目标,这类结构可以分为单级单目标、单级多目标和多级多目标等三种类型.多级多目标的递阶结构如图 1-3 所示.

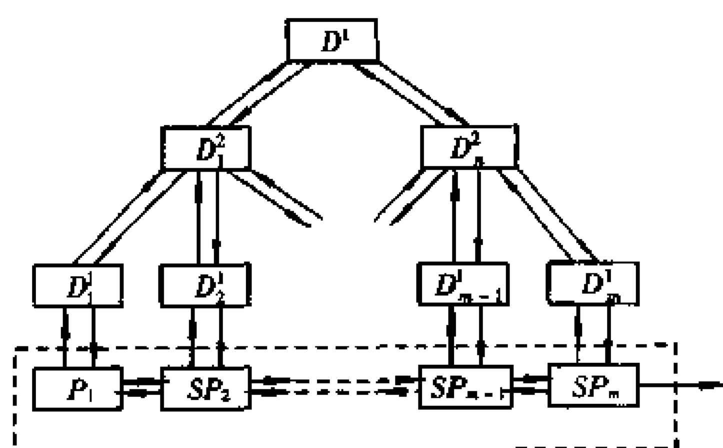


图 1-3

这类系统的决策单元处在不同的级别,按递阶排列,呈金字塔结构.只有上下级间才有信息交换,同级之间不交换信息.目标可能有冲突,它通过上一级的协调来解决.协调的最后结果应该就是或者近似于按全局优化的结果.

以上所述的三种大系统的基本递阶形式,是从不同角度考虑而形成的.多重描述主要是从建模来考虑的;多层描述是把一个复杂的决策问题进行纵向分解,按任务的复杂性分成若干子决策层;而多级描述则是考虑各子系统的关联,把决策问题进行横向分解.它们的共同特点归纳如下:

第一,越是处于高层的单元,对系统行为影响的范围也就越广.

第二,高级单元的决策周期,要比低级单元的决策周期长,这主要是因为要处理涉及到系统行为中变化较慢的因素.

第三,越是处于高级的问题,其问题的描述就会遇到更多的不确定性,因而更难于定量地予以公式化.

应该指出,三种形式的递阶结构不是互相排斥的,有时它们可同时存在于一个系统之中.

2　大系统的模型简化

由于大系统包含的元件多,元件之间关联复杂,输入和输出的数目也很多,建立大系统精确数学模型不仅困难,而且存在计算上的复杂性,因此,需要对大系统的数学模型进行简化.通常采用的简化方法有集结法和奇异摄动法.

2.1　集结法简化大系统模型

所谓集结法,就是把原系统中众多的状态变量按线性组合,合并成数量较少的新状态变量的方法.集结后的简化模型保持了原模型的主导特征值,使简化后模型的动态特性与原模型的动态特性无很大的差异.如原模型的阶次为 n,集结后简化模型的阶次为 $m, n > m$. m 的选择是集结法简化模型的关键,实质上也是一个系统阶次的辨识问题.1968 年阿奥基(M.Aoki)首先把集结法用于简化大规模动态系统.

对于大规模线性时不变系统,集结过程可表示如下:

考虑线性能控系统

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t),$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{2-1}$$

$$\boldsymbol{y}(t) = \boldsymbol{D}\boldsymbol{x}(t). \tag{2-2}$$

其中 $\boldsymbol{x}(t)$, $\boldsymbol{u}(t)$ 和 $\boldsymbol{y}(t)$ 分别为 $n\times 1$, $m\times 1$ 和 $r\times 1$ 的状态、控制和输出向量;$\boldsymbol{A}$, $\boldsymbol{B}$ 和 $\boldsymbol{D}$ 是 $n\times n$, $n\times m$ 和 $r\times n$ 的矩阵.要求上述系统的时间特性用下式进行描述:

$$z(t) = Cx(t), \quad z(0) = z_0 = Cx_0, \tag{2-3}$$

其中 C 是 $l\times n(l < n)$ 常数集结矩阵;z 是 $l\times 1$ 向量,称为 x 的集结.另一方面,假定 x 存在,并从 $z_0 = Cx_0$ 开始,希望保持(2-3)式的关系.不失一般性,假定秩 $\mathrm{rank}(C) = l$.这样,集结系统可表示为

$$\dot{z}(t) = Fz(t) + Gu(t),$$

$$z(0) = z_0, \tag{2-4a}$$

$$\hat{y} = Kz(t). \tag{2-4b}$$

其中$(\boldsymbol{F}, \boldsymbol{G})$对满足以下所谓动态精确性(理想集结)条件:

$$\boldsymbol{FC} = \boldsymbol{CA}, \tag{2-5}$$

$$\boldsymbol{G} = \boldsymbol{CB}, \tag{2-6a}$$

$$\boldsymbol{KC} \cong \boldsymbol{D}. \tag{2-6b}$$

向量 $\hat{y}$ 是一个 $r\times 1$ 的近似输出.注意到,如果 $l < n$,并假定系统(2-2)式是不可简

约的,则(2-5)式 ~ (2-6)式是不能同时满足的.因此,条件(2-6b)是近似的.

如果定义误差量为

$$e(t) = z(t) - Cx(t),$$

则它的动态特性由

$$\dot{e} = Fe(t) + (FC - CA)x(t) + (G - CB)u(t)$$

给定.如果条件(2-5)式 ~ (2-6)式满足,则 $\dot{e}(t) = Fe(t)$.因此,如果 $e(0) = 0$,则对于所有 $t \geqslant 0, e(t) = 0$.若 $e(0) \neq 0$,但 F 是稳定矩阵,则

$$\lim_{t \to \infty} e(t) = 0,$$

即动态精确性条件(2-5)式 ~ (2-6)式是渐近满足的.

确定集结矩阵(F,G),可用如下二种方法.

(1) 彭罗斯法

彭罗斯(Penrose)方法是根据彭罗斯可解性条件来求集结矩阵(F,G)的,即

$$F = CAC^{T}[CC^{T}]^{-1}. \tag{2-7}$$

一旦知道 C,受集结的矩阵 F 就可从(2-7)式得到,但它必须满足(2-5)式 ~ (2-6)式的条件.此时 F 的特征值构成了 A 的特征值的子集.这个方法的缺点就是要求有 A 的所有特征值的知识.

(2) 阿奥基方法

这种方法是由阿奥基提出的.

考虑能控矩阵

$$W_A \stackrel{\text{def}}{=\!=} [B, AB, \cdots, A^{n-1}B] \tag{2-8}$$

和修正后(2-4)式的能控矩阵

$$W_F \stackrel{\text{def}}{=\!=} [G, FG, \cdots, F^{n-1}G]. \tag{2-9}$$

由(2-5)式和(2-6b)式可知,上述二矩阵 W_A 和 W_F 满足

$$W_F = CW_A. \tag{2-10}$$

利用广义(伪)逆,矩阵 C 可以由下式获得

$$C = W_F W_A^{+} = W_F W_A^{T}(W_A W_A^{T})^{-1}. \tag{2-11}$$

因为由初始能控性的假设,秩 $\text{rank}(W_A) = n$.如果指定 F,例如 $F = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_l)$,选择 G 使(2-4)式完全能控,即秩 $\text{rank}(W_F) = l$,于是,C 可从(2-11)式求得.

2.2　摄动法简化大系统模型

用摄动法简化大系统模型是科科托维奇(P.V.Kokotovic)在 1972 年提出的.所谓摄动法是指在系统数学模型中用某些数量级较低的小参数的变动来模拟真实系统和近似系统响应差别的方法.当小参数摄动不致严重地改变动态特性时,称为正则摄动,它可以用来简化"弱耦合"的大系统.用小参数摄动方法研究系统某些特殊情况下的特征,称为奇异摄动.对"紧耦合"大系统就用奇异摄动法来简化.它把大系统动态过程中快变模和慢变模分开,先略去快变模,使系统简化,然后用"伸长

的时标”计算边界层校正,加入快变模的效应,以改进逼近度.

2.2.1 弱耦合模型

在许多工业控制系统中,可忽略某些动态耦合以减轻系统分析、设计的计算量.考虑以下分成 k 个子系统的大系统:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \vdots\\ \dot{x}_k\end{bmatrix}=\begin{bmatrix}A_1 & \varepsilon A_{12} & \cdots & \varepsilon A_{1k}\\ \varepsilon A_{21} & A_2 & \cdots & \varepsilon A_{2k}\\ \vdots & \vdots & & \vdots\\ \varepsilon A_{k1} & \varepsilon A_{k2} & \cdots & A_k\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{bmatrix}+\begin{bmatrix}B_1 & \varepsilon B_{12} & \cdots & \varepsilon B_{1k}\\ \varepsilon B_{21} & B_2 & \cdots & \varepsilon B_{2k}\\ \vdots & \vdots & & \vdots\\ \varepsilon B_{k1} & \varepsilon B_{k2} & \cdots & B_k\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_k\end{bmatrix}. \tag{2-12}$$

其中 ε 是一个小的正耦合参数;x_i 和 u_i 分别是第 i 个子系统的状态和控制向量;所有 A 和 B 矩阵假设为恒定.

对于(2-12)式系统,基于**正则摄动**建立的模型称为**弱耦合模型**.(2-12)式的特殊情况是当 $k=2$ 时,科科托维奇等称之为“ε 耦合”系统,即

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}=\begin{bmatrix}A_1 & \varepsilon A_{12}\\ \varepsilon A_{21} & A_2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}B_1 & \varepsilon B_{12}\\ \varepsilon B_{21} & B_2\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}. \tag{2-13}$$

显然,当 $\varepsilon=0$ 时,上述系统解耦而成二个子系统:

$$\dot{\hat{x}}_1=A_1\hat{x}_1+B_1\hat{u}_1,$$

$$\dot{\hat{x}}_2=A_2\hat{x}_2+B_2\hat{u}_2. \tag{2-14}$$

这就是二个近似的集结模型.它大大减轻了阶次 n 和 k 大于2的大系统的设计和分析的计算量.

2.2.2 强耦合模型

强耦合系统是指其变量具有差别很大的变化速度的系统.对于这种系统,基于“奇异摄动”建立的模型称为**强耦合模型**.它与弱耦合模型的区别在于摄动放在状态方程的左边,即小参数乘上状态变量的导数.

考虑由以下方程描述的奇异摄动系统:

$$\begin{cases}\dot{x}=A_1x(t)+A_{12}z(t)+B_1u(t),\\ x(t_0)=x_0,\end{cases} \tag{2-15}$$

$$\begin{cases}\varepsilon\dot{z}(t)=A_{21}x(t)+A_2z(t)+B_2u(t),\\ z(t_0)=z_0.\end{cases} \tag{2-16}$$

如果 A_2 是非奇异的,则当 $\varepsilon\to 0$ 时,(2-15)式和(2-16)式变成

$$\dot{\hat{x}}(t)=(A_1-A_{12}A_2^{-1}A_{21})\hat{x}+(B_1-A_{12}A_2^{-1}B_2)\hat{u}, \tag{2-17}$$

$$\hat{z}=-A_2^{-1}A_{21}\hat{x}-A_2^{-1}B_2\hat{u}. \tag{2-18}$$

方程(2-17)式是(2-15)式和(2-16)式的近似集结模型,实质上意味着原系统的 n

个特征值由$(A_1 - A_{12}A_2^{-1}A_{21})$矩阵的 l 个特征值来近似.

3　大系统递阶控制

递阶控制的概念是 20 世纪 70 年代,由米沙罗维奇(M.D.Mesarovic)等人提出来的.根据受控对象的性质,存在**稳态大系统递阶控制**和**动态大系统递阶控制**两种控制.前者适用于过程变化较慢的工业系统,它关心的是稳态工况的优劣,按给定的指标决定最优的稳态工作点;后者关心的是系统的动态品质,按给定的指标决定最优的状态轨迹.

递阶控制的基本思想是将大系统分解成若干相对独立的子系统,并构成控制系统的下层,而用上层的协调器来处理各子系统之间的关联作用.不论是稳态的还是动态的大系统,协调所采用的方法主要有关联预测法和关联平衡法等两种.

(1) **关联预测法**　关联预测法是协调器预测各子系统的关联输入和输出变量,下层各决策单元按预测的关联值求解各自的决策问题,然后把达到的性能指标送给协调器,协调器再修正预测值,直到总体目标达到最优为止.由于在协调过程中模型中引入了约束,这方法也称**模型协调法**.这个方法的中间结果是物理上可实现的,因此也叫**可行分解法**.

(2) **关联平衡法**　关联平衡法在下层各决策单元求解自己优化问题时,不考虑关联约束,协调器通过干预信息来修正各决策单元的优化目标,以保证最后关联约束得到满足.这时目标修正项的值也趋于零,达到原目标的最优值.此方法也称**目标协调法**.其中间结果不能施加于实际系统中,故也叫**不可行分解法**.

下面分别介绍稳态大系统和动态大系统的递阶优化控制.

3.1　稳态大系统的递阶控制

大工业生产过程都比较复杂,涉及到许多相互关联子系统,包含很多个控制输入变量、输出变量和内部的关联变量,这些变量还相互受到某些约束的限制.对于这种系统如果还是采用单一的控制器并按照常规的数学规划来决定整个系统的稳态最优工作点,就会遇到计算工作量和存储上的困难.20 世纪 70 年代初期米沙罗维奇等人提出的多级和递阶控制的概念为解决这样一种复杂大系统的优化提供了新途径.

3.1.1　稳态大系统递阶控制问题的提法

一个受控复杂大系统可以看成是若干(N)个子系统的关联组合,如图 3-1 所示。

对于第 i 个子系统,它的稳态输出向量 $\boldsymbol{y}_i^*$ 和控制向量 $\boldsymbol{c}_i$,关联输入向量 $\boldsymbol{u}_i^*$ 以及外界扰动向量 z_i 之间的关系一般可表示成:

$$\boldsymbol{y}_i^* = \boldsymbol{f}_i^*(\boldsymbol{c}_i, \boldsymbol{u}_i^*, z_i) \qquad (i = 1, 2, \cdots, N), \tag{3-1}$$

其中 $\boldsymbol{y}_i^* \in \mathbf{R}^{n_i}$;$\boldsymbol{c}_i \in \mathbf{R}^{m_i}$;$\boldsymbol{u}_i^* \in \mathbf{R}^{r_i}$;$z_i \in \mathbf{R}^{l_i}$;$\boldsymbol{f}_i^*$ 表示某一特定的 n_i 维向量函数关

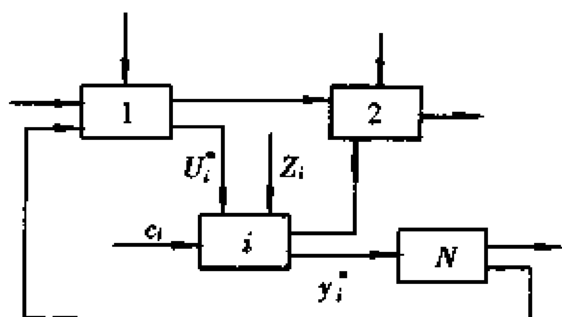


图 3-1

系，一般都是非线性的. 关联输入向量 u_i^* 来自其他子系统(也包括第 i 个子系统本身)的输出，它们之间的耦合约束关系为

$$u_i^* = \sum_{j=1}^{N} H_{ij} y_j^*, \tag{3-2}$$

其中 H_{ij} 是 $r_i \times n_j$ 维的布尔型矩阵，其元素由0或1组成，反映了从第 j 个子系统的输出到第 i 个子系统输入的关联.

如果设

$$c \xlongequal{\text{def}} (c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, \cdots, c_N^{\mathrm{T}})^{\mathrm{T}},$$
$$y^* \xlongequal{\text{def}} (y_1^{*\mathrm{T}}, y_2^{*\mathrm{T}}, \cdots, y_N^{*\mathrm{T}})^{\mathrm{T}},$$
$$u^* = (u_1^{*\mathrm{T}}, u_2^{*\mathrm{T}}, \cdots, u_N^{*\mathrm{T}})^{\mathrm{T}},$$
$$z = (z_1^{\mathrm{T}}, z_2^{\mathrm{T}}, \cdots, z_N^{\mathrm{T}})^{\mathrm{T}},$$
$$f^* = (f_1^{*\mathrm{T}}, f_2^{*\mathrm{T}}, \cdots, f_N^{*\mathrm{T}})^{\mathrm{T}},$$
$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1N} \\ \vdots & \vdots & & \vdots \\ H_{N1} & H_{N2} & \cdots & H_{NN} \end{bmatrix},$$

则整个系统的输入输出关系和关联约束关系为

$$\begin{aligned} y^* &= f^*(c, u^*, z), \\ u^* &= Hy^*. \end{aligned} \tag{3-3}$$

假设对于加到实际系统(3-3)式上的每一控制向量 c 和扰动向量 z，系统将产生唯一的输出 y^*. 这样整个系统从数学上又可看成是从 c 和 z 到 y^* 的单值映射 K^*：

$$\sum m_i \times \sum l_i \to \sum n_i,$$
$$y^* = K^*(c, z).$$

对于每一个子系统，都可以规定一个具体的控制性能指标 Q_i，它依 c_i，u_i^* 和 y_i^* 而定，记作 $Q_i(c_i, u_i^*, y_i^*)$. 它是一个标量，其大小从某些方面反映了生产效果的好坏. 另外从安全和系统中的物理性能以及实际条件的限制等方面来考虑，控制量 c_i、关联输入 u_i^* 和 y_i^* 之间，往往还要受到一定的约束. 这些约束条件在数学上可用一组等式或不等式所规定的集合 CUY_i 来表示，其一般形式是

$$
\begin{aligned}
&(c_i, u_i^*, y_i^*) \in CUY_i \xlongequal{\text{def}} \{(c_i, u_i, y_i)\},\\
&g_{ij}(c_i, u_i, y_i) \leqslant 0, j \in J_i \qquad (i = 1,2,\cdots,N),
\end{aligned}
\tag{3-4}
$$

其中 $g_{ij}, j = 1,2,\cdots,J_i$ 是一组线性或非线性的函数,其变元可以包括 c_i, u_i, y_i,也可仅含其中的一个或二个.所期望的稳态最优控制就是要选取所有子系统的控制输入向量 $c_i, i = 1,2,\cdots,N$,使得整个系统的性能指标

$$
Q = \sum_{i=1}^{N} Q_i(c_i, u_i^*, y_i^*)
$$

取极小值或极大值,同时又保证不违反各子系统所要求的约束(3-4)式.

要解决以上问题所遇到的困难是,实际上并不可能准确地知道各系统的特性(3-1)式,而只能用近似的数学模型来描述.另外,外界扰动 z,作为时间的函数也很难准确地估计,但只要其变化不十分频繁,都可以认为它在整个控制周期内可取某一平均常量,也就是说,它对系统的影响是固定的,因而也就可以从模型中略去.这样就可用

$$
\begin{aligned}
y_i &= f_i(c_i, u_i),\\
u_i &= \sum_{j=1}^{N} H_{ij} y_j = H_i y \quad (i = 1,2,\cdots,N),\\
H_i &= (H_{i1}, H_{i2}, \cdots, H_{iN}),
\end{aligned}
\tag{3-5}
$$

近似描述各子系统的特性.其中 y_i 和 u_i 分别是模型的输出变量和关联输入变量.

对于整个系统,写成集合向量的形式,有

$$
\begin{cases}
y = f(c, u),\\
u = Hy,
\end{cases}
\tag{3-6}
$$

和

$$
y = K(c) \tag{3-7}
$$

按照模型(3-5)式求下述问题的解:

$$
\begin{aligned}
&\min_{c,u,y} Q(c, u, y),\\
&\text{s.t. } y = f(c, u),\\
&\qquad u = Hy,\\
&\qquad (c, u, y) \in CUY,
\end{aligned}
$$

得到的即是开环的最优解,它可以离线地求出.但由于 f_i 和 f_i^* 的差异,实际系统的输出 y_i^* 和 y_i 并不相等,因此,不能保证系统运行于实际的最优点,甚至连约束(3-4)式有时也不一定能够满足.于是,又提出了用模型和反馈信息来决定真实系统次优控制的方法,这就是所谓的在线(闭环)递阶控制算法.

3.1.2 关联预测法

关联预测法又叫直接协调法或模型协调法.它的基本思想是用指定子系统模型关联输出变量 $y_i, i = 1,2,\cdots,N$(同时也就规定了关联输入变量 u_i)的办法将各子系统去耦.

如用数学语言来具体描述下级各决策单元和上级协调器的任务，可以写成：

对于第 i 个局部决策单元，其任务为

$$LP_i \atop (i=1,2,\cdots,N) \begin{cases} \text{对于协调器指定的 } y \in Y\text{，求出控制向量 } \widehat{c}_i\text{，使得} \\ \widehat{c}_i = \arg\min_{c_i} Q_i(c_i, H_i y, y_i), \\ \text{同时满足} \quad y_i = f_i(c_i, H_i y), \\ \qquad (c_i, H_i y, y_i) \in CUY_i, \end{cases} \tag{3-8}$$

其中集合

$$CUY_i \xlongequal{\text{def}} \{(c_i, u_i, y_i) \mid g_{ij}(c_i, u_i, y_i) \leqslant 0, j = 1,2,\cdots,J_i\}.$$

集合 Y 定义为当 $y \in Y$ 时，集合 $c_i(y) \neq \varnothing, i = 1,2,\cdots,N$. 而 $c_i(y)$ 是指满足 $y_i = f_i(c_i, H_i y)$ 和 $(c_i, H_i y, y_i) \in CUY_i$ 的所有 c_i 的集合.

显然各局部决策单元求出的 $\widehat{c}_i$ 都是协调变量 y 的函数. 在求出 $\widehat{c}_i(y)$ 之后，各局部决策单元把相应的性能指标值 $Q_i(\widehat{c}_i(y), H_i y, y_i)$ 送给上级协调器，协调器的任务是

$$CP \begin{cases} \text{找到 } \widehat{y} \in Y \text{ 使得} \\ \sum_{i=1}^{N} Q_i(\widehat{c}_i(\widehat{y}), H_i\widehat{y}, \widehat{y}_i) \\ = \min_{y \in Y} \sum_{i=1}^{N} Q_i(\widehat{c}_i(y), H_i y, y_i), \end{cases} \tag{3-9}$$

这里集合 Y 的定义同前.

$\widehat{c}_i(\widehat{y}), i = 1,2,\cdots,N$，就是整个系统基于模型的最优控制向量. 在这组控制向量的作用下，模型的关联输出向量和输入向量分别为 $\widehat{y}$ 和 $H\widehat{y}$，它们之间满足模型方程和约束条件：

$$\widehat{y}_i = f_i(\widehat{c}_i(\widehat{y}), H_i\widehat{y}),$$

$$\widehat{c}_i(\widehat{y}, H_i\widehat{y}, \widehat{y}_i) \in CUY_i \quad (i = 1,2,\cdots,N), \tag{3-10}$$

整个问题的求解是通过逐次校正协调变量，并重复求解各局部决策单元的优化问题这样一个迭代过程来实现的. 协调器按照所选定的某种寻优程序进行搜索，协调变量的设定值每改变一次，各局部决策单元就进行一次优化计算，并把求得的最优性能指标值告诉协调器，以便为下一步搜索提供依据. 这个过程一直继续到求得最优的总体性能指标值为止.

关联预测法的一个重要优点是，在协调器寻优过程中的每一步，各子系统求得的 $\widehat{c}_i(y), i = 1,2,\cdots,N$，虽然还不是整个系统最优的控制向量，但把它加到系统上，模型方程和约束条件都能满足，因此，都是可行的，正因为有这个特点，所以这种方法也称为可行协调法.

3.1.3 关联平衡法

关联平衡法又称为目标协调法或价格协调法.它的基本思想是,割断各子系统之间的耦合,各局部决策单元把关联输入变量当做独立寻优变量来处理.而实际存在的关联约束式

$$u_i = \sum_{j=1}^{N} H_{ij} y_j \quad (i = 1,2,\cdots,N),$$

通过引入拉格朗日(J.L.Lagrange)乘子向量 $\lambda_i, i = 1,2,\cdots,N$,将其归并到系统的性能指标中去作为对目标的修正.一般对于任意给定的一组 $\lambda_i, i = 1,2,\cdots,N$,各局部决策单元所求得的使修正性能指标为极小的 $\hat{c}(\lambda)$ 和 $\hat{u}_i(\lambda)$ 并不满足关联约束条件:

$$u_i = \sum_{j=1}^{N} H_{ij} y_j,$$

其中 $y_j = f_j(c_j, u_j)$.

协调器的任务则是,通过选择适当的拉格朗日乘子(协调变量)来协调各子系统的控制性能指标,以保证最后关联约束条件得以满足,也就是达到关联平衡.

考虑关联约束后,整个系统的拉格朗日函数为

$$L(c,u,y,\lambda) = \sum_{i=1}^{N} \{Q_i(c_i,u_i,y_i) + \lambda_i^{\mathrm{T}}(u_i - \sum_{j=1}^{N} H_{ij} y_j)\},$$

它可以分解成 N 个拉格朗日函数之和

$$L = \sum_{i=1}^{N} L_i,$$

$$L_i = Q_i(c_i,u_i,y_i) + \lambda_i^{\mathrm{T}} u_i - \sum_{j=1}^{N} \lambda_j^{\mathrm{T}} H_{ji} y_i.$$

其中第 i 个拉格朗日函数只与第 i 个子系统的控制向量 c_i,关联输入向量 u_i 和关联输出向量 y_i 有关;$\lambda_j, j = 1,2,\cdots,N$,是由协调器指定的;$L_i$ 即是修正后的第 i 个子系统的性能指标.显然,当最终达到关联平衡时,对性能指标的修正也就消失了.这样,第 i 个局部决策单元所需解决的问题为

$$\begin{matrix} LP_i \\ (i = 1,2,\cdots,N) \end{matrix} \left\{ \begin{array}{l} \text{对于协调器指定的 } \lambda = (\lambda_1,\lambda_2,\cdots,\lambda_N), \\ \text{求出 } \hat{c}_i(\lambda) \text{ 和 } \hat{u}_i(\lambda)\text{,使得} \\ \quad (\hat{c}_i(\lambda),\hat{u}_i(\lambda)) = \arg \min\limits_{c_i,u_i} L_i(c_i,u_i,y_i,\lambda), \\ \quad \text{s.t. } y_i = f_i(c_i,u_i), \\ \quad (c_i,u_i,y_i) \in CUY_i, \end{array} \right. \quad (3\text{-}11)$$

其中

$$L_i(c_i,u_i,y_i,\lambda) = Q_i(c_i,u_i,y_i) + \lambda_i^{\mathrm{T}} u_i - \sum_{j=1}^{N} \lambda_j^{\mathrm{T}} H_{ji} y_i,$$

$$CUY_i \stackrel{\text{def}}{=\!=} \{(c_i,u_i,y_i) \mid g_{ij}(c_i,u_i,y_i) \leqslant 0 \quad (j = 1,2,\cdots,J_i)\}.$$

显然 $\widehat{c}_i,\widehat{u}_i$ 都是 λ 的函数.

协调器的任务为

$$
CP\begin{cases}\text{求 } \widehat{\lambda}=(\widehat{\lambda}_1,\widehat{\lambda}_2,\cdots,\widehat{\lambda}_N),\text{使得}\\ \widehat{u}_i(\widehat{\lambda})-\sum_{j=1}^{n}H_{ij}\widehat{y}_j(\widehat{\lambda})=0 \quad (i=1,2,\cdots,N),\end{cases}\tag{3-12}
$$

其中 $\widehat{y}_i(\widehat{\lambda})=f_i(\widehat{c}_i(\widehat{\lambda}),\widehat{u}_i(\widehat{\lambda}))$.

如果定义对偶函数为

$$
D(\lambda)=\sum_{i=1}^{N}L_i(\widehat{c}_i(\lambda),\widehat{u}_i(\lambda),f_i(\widehat{c}_i(\lambda),\widehat{u}_i(\lambda)),\lambda),
$$

并注意到 $D(\lambda)$ 对 λ_i 的导数为

$$
\dot{D}_i(\lambda)=\widehat{u}_i(\lambda)-\sum_{j=1}^{N}H_{ij}\widehat{y}_j(\lambda),
$$

$$
\widehat{y}_i(\lambda)=f_i(\widehat{c}_i(\lambda),u_i(\lambda)),
$$

那么这一协调器问题就等效于寻找对偶函数 $D(\lambda)$ 梯度为零的点.根据拉格朗日对偶原理可知,如果原问题的拉格朗日函数具有鞍点,那么鞍点就是 $D(\lambda)$ 的极大值点,而且 $D(\lambda)$ 的极大值就等于原问题的极小值,即

$$
\begin{aligned}
&\max_{\lambda}D(\lambda)=\min_{c,u,y}Q(c,u,y),\\
&\text{s.t.}\quad y=f(c,y),\\
&\qquad u=Hy,\\
&\qquad (c,u,y)\in CUY.
\end{aligned}\tag{3-13}
$$

所以协调器可采用十分简单的梯度法来校正 λ,使对偶函数达到最大. $D(\lambda)$ 对 λ 的梯度即为关联平衡失调向量 $\widehat{u}(\lambda)-Hf(\widehat{c}(\lambda),\widehat{u}(\lambda))$. 当 $D(\lambda)$ 取极大值时,梯度为零,也就是实现了关联平衡.

与关联预测法相比,关联平衡法具有以下的特点:

(1) 各子系统决策单元的寻优变量数比用关联预测法要多,除了控制向量外,还有关联输入向量.

(2) 协调变量的个数等于关联约束条件的个数.由于可以采用梯度法,因此,协调算法比较简单.最早用这种方法来解决经济学中的一些问题时,需要处理的常常是供求平衡这样一些关联约束,这时协调变量 λ 有着明显的实际意义,就像价格一样起着调节供求平衡的作用,所以这种方法也叫价格协调法.

(3) 一般对于各子系统,不要求控制向量的维数一定要大于关联输出向量的维数,协调变量 λ 可在整个实数范围内取值.

(4) 这种方法的一个重要缺点是,迭代过程的任一中间结果 $\widehat{c}(\lambda),\widehat{u}(\lambda),\widehat{y}(\lambda)$ 只满足模型方程和约束集合,即满足

$$
\widehat{y}(\lambda)=f(\widehat{c}(\lambda),\widehat{u}(\lambda)),
$$

和 $$(\hat{c}(\lambda), \hat{u}(\lambda), \hat{y}(\lambda)) \in CUY,$$

而不满足关联约束条件 $\hat{u}(\lambda) = Hf(\hat{c}(\lambda), \hat{u}(\lambda))$，因此，其解是不可行的. 只有到迭代终止，找到 $\hat{\lambda}$后，关联约束才得以满足. 这时 $\hat{c}(\hat{\lambda})$ 加于实际系统才是安全的，所以这种方法是不可行的协调法.

3.1.4 混合法

在混合法中，协调变量不仅有拉格朗日乘子，还有各子系统之间的关联变量，它也是一种不可行方法. 其基本思想及方法兼有关联预测法和关联平衡法的特点.

3.2 动态线性大系统的递阶控制

3.2.1 动态线性大系统的互联数学模型

假定，如图 3-2 所示的**互联数学模型**，整个系统由互相连接的 N 个子系统组成. 对于任何子系统 i，x_i 是 n_i 维的状态向量，u_i 是 m_i 维的控制向量，z_i 是由其他子系统的状态产生的 r_i 维输入向量. 各子系统本身可以由线性微分方程来描述，即

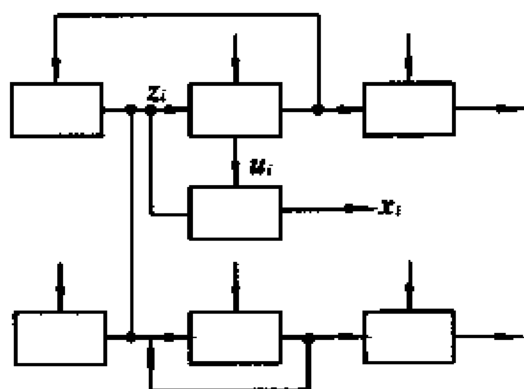


图 3-2

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + C_i z_i(t), \tag{3-14}$$
$$x_i(0) = x_{i0},$$

再进一步假定输入向量 z_i 是 N 个子系统状态的线性组合，即

$$z_i = \sum_{j=1}^{N} L_{ij} x_j, \tag{3-15}$$

大系统递阶优化问题就是选择控制 $u_1, u_2, \cdots, u_N$，使下述形式的二次性能指标函数为最小：

$$J = \sum_{i=1}^{N} \left(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \int_0^T \frac{1}{2} (\| x_i(t) \|_{Q_i}^2 + \| u_i \|_{R_i}^2 + \| z_i \|_{S_i}^2) dt \right). \tag{3-16}$$

其中 Q_i、p_i 是正半定矩阵；R_i，S_i 是正定矩阵；$\| z_i \|_{S_i}^2$ 很难给以一定的物理意义，但加上这一项，可避免求解过程中的奇异问题.(3-16)式受(3-14)式和(3-15)式的约束.

3.2.2　目标协调法

用目标协调法(关联平衡法)求解动态线性大系统优化控制的主要思想是，把原始的最小化问题转化成较简单的最大化问题，然后用二级迭代计算结构来求解.为此，需定义一个对偶函数 $\Phi(\lambda)$，这里

$$\Phi(\lambda) = \min_{x,u,z} L(x,u,z,\lambda) \tag{3-17}$$

服从于(3-1)式的约束.其中

$$L(x,u,z,\lambda) = \sum_{i=1}^{N}\Big(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \int_0^T (\frac{1}{2} \| x_i \|_{Q_i}^2 + \frac{1}{2} \| u_i \|_{R_i}^2 + \frac{1}{2} \| z_i \|_{S_i}^2 + \lambda_i^{\mathrm{T}}(z_i - \sum_{j=1}^{N} L_{ij}x_j(t)))\mathrm{d}t\Big), \tag{3-18}$$

其中 λ 和 λ_i 分别是 r 维和 r_i 维拉格朗日乘子向量，$r = \sum_{i=1}^{N} r_i$，$L(x,u,z,\lambda)$ 是通过引入拉格朗日乘子而组成的拉格朗日算式.根据拉格朗日对偶定理，像现在所有约束为线性且性能指标函数为二次型的情况，有

$$\max_{\lambda} \Phi(\lambda) = \min_{u} J, \tag{3-19}$$

也就是说，受(3-14)式和(3-15)式约束，求(3-16)式的极小问题，等效于对偶函数对 λ 求极大值的问题.

对于这种问题，可以给定 $\lambda = \lambda^*$，于是(3-18)式可写为

$$\begin{aligned} L(x,u,z,\lambda^*) &= \sum_{i=1}^{N}\Big(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \int_0^T (\frac{1}{2} \| x_i \|_{Q_i}^2 + \frac{1}{2} \| u_i \|_{R_i}^2 + \frac{1}{2} \| z_i \|_{S_i}^2 + \lambda_i^{*\mathrm{T}}(z_i - \sum_{j=1}^{N} \lambda_j^{*\mathrm{T}} L_{ji} x_i(t)))\mathrm{d}t\Big) \\ &= \sum_{i=1}^{N} L_i. \end{aligned} \tag{3-20}$$

这就是说，拉格朗日算式可以分解成 N 个独立的子拉格朗日算式.对于每一个子系统，拉格朗日乘子是已知的.在(3-14)式的约束下，使(3-20)式最小，其中 λ_i^* 可以作为已知函数由递阶结构的第二级给定.每个子系统的最小化结果可决定(3-17)式的对偶函数 $\Phi(\lambda^*)$.在第二级中所有子系统的解为已知，$\Phi(\lambda^*)$ 可以用典型的无约束的优化方法，譬如牛顿法、梯度法或共轭梯度法来改善.用梯度法的理由是因为 $\Phi(\lambda)$ 的梯度就是子系统的关联误差，即

$$\nabla_{\lambda_i} \Phi(\lambda) |_{\lambda=\lambda^*} = z_i - \sum_{j=1}^{N} L_{ij}x_j = e_i \quad (i = 1,2,\cdots,N). \tag{3-21}$$

这样，可以得到如图3-3所示那样的二级递阶算法结构.图3-3中第一级，在第二级

供给 $\lambda = \lambda^*$ 之后，使受子系统动态约束的 L_i 最小，其所得的结果 x_i, z_i 送回第二级. 在第二级，将这些向量集合起来，代入(3-21)式，得到关联误差.

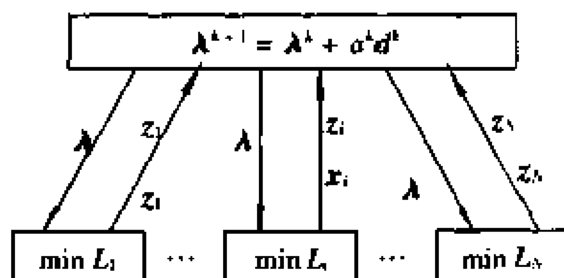


图 3-3

这个误差，在梯度法中用来产生新的 λ. 例如，从 k 次迭代到 $k+1$ 次迭代：

$$\lambda^{k+1}(t) = \lambda^k(t) + \alpha^k d^k(t) \quad (0 \leqslant t \leqslant T), \tag{3-22}$$

其中 α^k 是步长；d^k 是搜索方向. 从计算观点来看，共轭梯度法是比较好的，在这种情况下，有

$$d^{k+1}(t) = e^{k+1}(t) + \beta^{k+1} d^k(t) \quad (0 \leqslant t \leqslant T), \tag{3-23}$$

其中

$$\beta^{k+1} = \frac{\int_0^T (e^{k+1}(t))^{\mathrm{T}} e^{k+1}(t)\mathrm{d}t}{\int_0^T (e^k(t))^{\mathrm{T}} e^k(t)\mathrm{d}t}, \tag{3-24}$$

且

$$d^0 = e^0.$$

一旦 $e^k(t)(0 \leqslant t \leqslant T)$ 足够地接近于零，就达到全局最优.

这个方法又称为关联平衡法，因为在最优时，梯度为零，关联达到平衡：

$$z_i = \sum_{j=1}^{N} L_{ij} x_j.$$

由于在迭代过程中，各子系统的目标将不断地修正，所以一开始称它为目标协调法. 这个方法的主要缺点是，在第二级中数值计算收敛慢，它反过来又会影响到第一级的计算；它不能用较少的迭代次数来获得次优解，这是因为在第二级中未收敛的次优解是不可行的，即不到收敛终点不能满足以下条件：

$$z_i = \sum_{j=1}^{N} L_{ij} x_j.$$

3.2.3　关联预测法

另一个用在闭环和开环的递阶控制方法是关联预测法. 这种方法可避免在第二级中使用梯度型的迭代.

设整个系统由 N 个互相连接的线性子系统组成：

$$\dot{x}_i = A_i x_i + B_i u_i + C_i z_i \quad (i = 1, 2, \cdots, N), \tag{3-25}$$

其中

$$z_i = \sum_{j=1}^{N} L_{ij} x_j. \tag{3-26}$$

其目标是要使以下的性能指标最小：

$$J = \sum_{i=1}^{N} \left(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \frac{1}{2} \int_0^T (\| x_i(t) \|_{Q_i}^2 + \| u_i(t) \|_{R_i}^2) dt \right). \tag{3-27}$$

这里要注意，在二次项中没有 z_i 这一项，其拉格朗日函数为

$$L = \sum_{i=1}^{N} \left(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \int_0^T (\frac{1}{2} \| x_i(t) \|_{Q_i}^2 + \frac{1}{2} \| u_i(t) \|_{R_i}^2 + \lambda_i^T (z_i - \sum_{j=1}^{N} L_{ij} x_j) + p_i^T (- \dot{x}_i + A_i x_i + B_i u_i + C_i z_i)) dt \right), \tag{3-28}$$

其中 p_i 是 n_i 维的伴随向量，λ_i 是 r_i 维拉格朗日乘子向量. 对于给定的 $\lambda_i = \lambda_i^*$，$z_i = z_i^*$，(3-28) 式就成为可分的了，即

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left(\frac{1}{2} \| x_i(T) \|_{P_i}^2 + \int_0^T (\frac{1}{2} \| x_i(t) \|_{Q_i}^2 + \frac{1}{2} \| u_i(t) \|_{R_i}^2 + \lambda_i^{*T} z_i^* - \sum_{j=1}^{N} \lambda_j^{*T} L_{ji} x_i + p_i^T (- \dot{x}_i + A_i x_i + B_i u_i + C_i z_i^*)) dt \right). \tag{3-29}$$

这里，不像目标协调法那样其协调向量只是拉格朗日乘子；现在协调向量是$\begin{bmatrix} \lambda \\ z \end{bmatrix}$，这比目标协调法的协调向量的阶次要高，然而在第二级中算法却极其简单，所以用较复杂的协调向量并没有什么缺点，只要用本次迭代所得的向量值当作下次迭代的向量值即可，即

$$\begin{pmatrix} \lambda^{*k+1} \\ z^{*k+1} \end{pmatrix} = \begin{pmatrix} \lambda^* (x^k, u^k, p^k) \\ z^* (x^k, u^k, p^k) \end{pmatrix}, \tag{3-30}$$

而等式的右边可以用驻点的条件获得，即

$$\frac{\partial L}{\partial z_i^*} = 0,$$

$$\lambda_i^* = - C_i^T p_i,$$

所以在第 $k+1$ 次的迭代中，第二级产生向量

$$\begin{pmatrix} \lambda^{*k+1} \\ z^{*k+1} \end{pmatrix} = \begin{pmatrix} - C_i^T p_i \\ \sum_{j=1}^{N} L_{ij} x_j \end{pmatrix}^k. \tag{3-31}$$

在第一级，在给定 λ_i^* 和 z_i^* 的情况下，子系统只是解一个标准的两点边值问题. 对于第 i 个子系统，其哈密顿(Hamilton) 函数可以定义为

$$H_i = \frac{1}{2} x_i^T Q_i x_i + \frac{1}{2} u_i^T R_i u_i + \lambda_i^T z_i - \sum_{j=1}^{N} \lambda_j^T L_{ji} x_i +$$

$$p_i^{\mathrm{T}}[A_ix_i + B_iu_i + C_iz_i]. \tag{3-32}$$

优化的必要条件可以写成

$$\dot{p}_i = -\frac{\partial H_i}{\partial x_i} = -Q_ix_i - A_{ii}^{\mathrm{T}}p_i + \sum_{j=1}^{N} L_{ji}^{\mathrm{T}}\lambda_j(t), \tag{3-33}$$

$$p_i(T) = \frac{\partial\left(\frac{1}{2}x_i^{\mathrm{T}}(T)P_ix_i(T)\right)}{\partial x_i(T)} = P_ix_i(T), \tag{3-34}$$

$$\begin{cases} \dot{x}_i(t) = \dfrac{\partial H_i}{\partial p_i} = A_ix_i(t) + B_iu_i(t) + C_iz_i(t), \\ x_i(0) = x_{i0}, \end{cases} \tag{3-35}$$

$$\frac{\partial H_i}{\partial u_i} = R_iu_i(t) + B_{ii}^{\mathrm{T}}p_i(t) = 0. \tag{3-36}$$

在第一级中，这里的 $\lambda_i(t)$ 和 $z_i(t)$ 不再是未知的.为了解第一级问题，从(3-36)式，得

$$u_i(t) = -R_i^{-1}B_{ii}^{\mathrm{T}}p_i(t), \tag{3-37}$$

将上式代入(3-33)式～(3-35)式，得

$$\dot{x}(t) = A_ix_i(t) - S_ip_i(t) + C_iz_i(t),$$

$$x_i(0) = x_{i0},$$

$$\dot{p}_i(t) = -Q_ix_i(t) - A_{ii}^{\mathrm{T}}p_i(t) + \sum_{j=1}^{N} L_{ji}^{\mathrm{T}}\lambda_j(t),$$

$$p_i(T) = P_ix_i(T).$$

这就是一个线性两点边值问题，且 $S_i = B_iR_i^{-1}B_i^{\mathrm{T}}$.引入矩阵里卡蒂(J.F.Riccati)方程，这问题可以去耦.假定

$$p_i(t) = K_i(t)x_i(t) + g_i(t), \tag{3-38}$$

其中 $g_i(t)$ 是一个 n_i 维开环"伴随"与"补偿"向量.将(3-38)式两边微分，并把(3-33)式和(3-35)式代入其中，经过演算就可以得到以下的矩阵和向量微分方程：

$$\dot{K}_i(t) = -K_i(t)A_i - A_i^{\mathrm{T}}K_i(t) + K_i(t)S_iK_i(t) - Q_i, \tag{3-39}$$

$$\dot{g}_i(t) = -[A_i - S_iK_i(t)]^{\mathrm{T}}g_i(t) - K_i(t)C_iz_i(t) + \sum_{j=1}^{N} L_{ji}^{\mathrm{T}}\lambda_j(t), \tag{3-40}$$

从(3-34)式和(3-38)式，求终值条件 $K_i(T)$ 和 $g_i(T)$，可得

$$K_i(T) = P_i,$$

$$g_i(T) = 0. \tag{3-41}$$

于是第一级控制规律变为

$$u_i(t) = -R_i^{-1}B_i^{\mathrm{T}}K_i(t)x_i(t) - R_i^{-1}B_i^{\mathrm{T}}g_i(t),$$

其中具有一个部分反馈(闭环)项和一个前馈(开环)项.

这里要注意两点：

第一，微分矩阵里卡蒂方程的解涉及 $n_i(n_i + 1)/2$ 个非线性标量方程，它与初始条件 $x_i(0)$ 无关；

第二,在(3-40)式中,$z_i(t)$ 和 $K_i(t)$、$g_i(t)$ 不同,它实质上与初始条件 $x_i(0)$ 有关.

关联预测法在第二级的计算要比目标协调法简单得多,而且不存在奇异问题.计算表明,关联预测法在第二级的收敛是相当快的.

3.2.4 离散大系统的三级协调法

离散大系统的三级协调法首先是由塔姆拉(J.Tamura)提出来的,因此又叫塔姆拉方法,它是建立在目标协调的基础上的.

要解的问题是使下式特性指标最小:

$$J=\sum_{i=1}^{N}\left(\frac{1}{2}\|x_i(k)\|^2_{G_i(k)}+\sum_{k=0}^{K-1}\frac{1}{2}\left(\|x_i(k)\|^2_{Q_i(k)}+\|z_i(k)\|^2_{S_i(k)}+\|u_i(k)\|^2_{R_i(k)}\right)\right)$$

其中$\frac{1}{2}\|x_i(k)\|^2_{G_i(k)}$ 是终点的指标;$\sum$ 里边的各项表示优化序列其余部分的指标,即 k 从 0 到 $K-1$.

最小化过程必须受子系统的动态约束,即

$$x_i(k+1)=A_ix_i(k)+B_iu_i(k)+C_iz_i(k) \tag{3-42}$$

$$(i=1,2,\cdots,N;k=0,1,2,\cdots,K-1),$$

假定初始条件已知,即

$$x_i(0)=x_{i0}, \tag{3-43}$$

像在连续系统一样,z_i 是关联输入向量,来自其他子系统,即

$$z_i(k)=\sum_{j=1}^{N}L_{ij}x_j(k) \tag{3-44}$$

$$(k=0,1,2,\cdots,K-1;i=1,2,\cdots,N),$$

为了解这个问题,像连续系统一样,要对拉格朗日乘子 λ 求对偶函数 $\Phi(\lambda)$ 的最大值,这里

$$\Phi(\lambda)=\min_{x,u,z}L(x,u,z,\lambda), \tag{3-45}$$

它受(3-42)式和(3-43)式的约束.构造拉格朗日函数

$$L(x,u,z,\lambda)=\sum_{i=1}^{N}\left\{\frac{1}{2}\|x_i(k)\|^2_{G_i(k)}+\sum_{k=0}^{K-1}\left(\frac{1}{2}\|x_i(k)\|^2_{Q_i(k)}+\frac{1}{2}\|z_i(k)\|^2_{S_i(k)}+\frac{1}{2}\|u_i(k)\|^2_{R_i(k)}+\lambda_i^{\mathrm{T}}z_i(k)-\sum_{j=1}^{N}\lambda_j^{\mathrm{T}}L_{ji}x_i(k)\right)\right\}=\sum_{i=1}^{N}L_i, \tag{3-46}$$

其中

$$L_i=\frac{1}{2}\|x_i(k)\|^2_{G_i(k)}+\sum_{k=0}^{K-1}\left(\frac{1}{2}\|x_i(k)\|^2_{Q_i(k)}+\frac{1}{2}\|u_i(k)\|^2_{R_i(k)}+\frac{1}{2}\|z_i(k)\|^2_{S_i(k)}+\right.$$

$$\lambda_i^{\mathrm{T}} z_i(k) - \sum_{j=1}^{N} \lambda_j^{\mathrm{T}} L_{ji} x_i(k) \Big) . \tag{3-47}$$

这样，像连续系统一样，对于由第二级给定的序列 $\lambda = \lambda^*$，有可能将拉格朗日 L 分解成 N 个独立的、受(3-42)式和(3-43)式约束的子拉格朗日函数 L_i，拉格朗日乘子向量序列可以用梯度型算法在第二级得到改善：

$$\nabla_{\lambda_i} \Phi(\lambda) \Big|_{\lambda=\lambda^*} = z_i(k) - \sum_{j=1}^{N} L_{ij} x_j(k) - e_i(k) \tag{3-48}$$

$$(i = 1,2,\cdots,N, k = 0,1,2,\cdots,K-1).$$

塔姆拉三级协调方法的基本出发点是两级结构的"第一级"的解可以用对偶原理进一步分解，而不去解 N 个独立的子问题(受(3-42)式和(3-43)式限制，使 L_i 最小化)，即在这一级中，对每一个子系统的子拉格朗日函数，可用离散时间指数进一步分解．这样将一个二级结构中第一级的"函数"优化问题变成三级结构中第一级的"参数"优化问题．这种分解是按时间进行的，而不是像前面讲的按子系统进行的．

为了确定这个三级结构的最优策略，定义受(3-42)式约束，使(3-47)式最小化的对偶问题为

$$\max_{p_i} M(p_i) \quad (i = 1,2,\cdots,N), \tag{3-49}$$

其中

$$\begin{aligned} M(p_i) = \min_{x_i,u_i,z_i} \Big\{ & \frac{1}{2} \| x_i(k) \|^2_{G_i(k)} + \sum_{k=0}^{K-1} \Big(\frac{1}{2} \| x_i(k) \|^2_{Q_i(k)} + \\ & \frac{1}{2} \| u_i(k) \|^2_{R_i(k)} + \frac{1}{2} \| z_i(k) \|^2_{S_i(k)} + \\ & p_i^{\mathrm{T}}(k)(A_i x_i(k) + B_i u_i(k) + C_i z_i(k) - x_i(k+1)) + \\ & \lambda_i^{*\mathrm{T}} z_i - \sum_{j=1}^{N} \lambda_j^{*\mathrm{T}} L_{ji} x_i \Big) \Big\} \quad (i = 1,2,\cdots,N) \end{aligned} \tag{3-50}$$

服从于已知初始状态 $x_i(0) = x_{i0}$．

为了在数值上解决这个对偶问题，必须用给定的 $p_i = p_i^*$，来计算对偶函数 $M(p_i)$，然后用某些梯度技术来对 p_i 求 $M(p_i)$ 的极大值，$M(p_i)$ 的梯度由下式给出：

$$\nabla_{p_i} M(p_i) |_{p_i = p_i^*} = - x_i(k+1) + A_i x_i(k) + B_i u_i(k) + C_i z_i(k) \tag{3-51}$$

$$(k = 0,1,2,\cdots,K-1, \quad i = 1,2,\cdots,N),$$

其中 x_i, u_i 是对于给定的 $p_i = p_i^*$，受(3-42)式约束，使(3-47)式的 L_i 最小化所得到的结果．

现在来定义第 i 个子系统的哈密顿函数 $H_i(\cdot)$．不失一般性，假定矩阵 G_i, Q_i, R_i, S_i 为常数，有

$$H_i(x_i(k), u_i(k), z_i(k), k) = \frac{1}{2} \| x_i(k) \|^2_{Q_i} + \frac{1}{2} \| u_i(k) \|^2_{R_i} +$$

$$\frac{1}{2}\|z_i(k)\|_{S_i}^2+\lambda_i^{*\mathrm{T}}z_i(k)-\sum_{j=1}^{N}\lambda_j^{*\mathrm{T}}L_{ji}x_i(k)+$$
$$p_i^{*\mathrm{T}}(k)[A_ix_i(k)+B_iu_i(k)+C_iz_i(k)]$$
$$(k=0,1,2,\cdots,K-1;\quad i=1,2,\cdots,N),\tag{3-52}$$

利用(3-50)式并考虑到

$$\sum_{k=0}^{K-1}p_i^{*\mathrm{T}}(k)x_i(k+1)-\sum_{k=0}^{K-1}p_i^{*\mathrm{T}}(k-1)x_i(k)$$
$$=p_i^{*\mathrm{T}}(K-1)x_i(K),\tag{3-53}$$

则 $M(p_i)$ 可以表达为

$$M(p_i)=\min_{x_i,u_i,z_i}\Big[\frac{1}{2}\|x_i(k)\|_{G_i}^2-p_i^{*\mathrm{T}}(K-1)x_i(K)+$$
$$\sum_{k=0}^{K-1}\Big(H_i(x_i(k),u_i(k),z_i(k),k)-p_i^{*\mathrm{T}}(k-1)x_i(k)\Big)\Big],\tag{3-54}$$

其中 $p_i(-1)$ 定义为零.所以对于给定的 $p_i=p_i^*$,最小化问题变为

(1) 对于 $k=0$,

$$\min_{u_i(0),z_i(0)}H_i(x_i(0),u_i(0),z_i(0),0),\tag{3-55}$$

服从于

$$x_i(0)=x_{i0},\tag{3-56}$$

上述问题的最小化的必要条件为

$$\nabla_{u_i(0)}H_i(\cdot)=R_iu_i(0)+B_i^{\mathrm{T}}p_i^*(0)=0,$$
$$\nabla_{z_i(0)}H_i(\cdot)=S_iz_i(0)+C_i^{\mathrm{T}}p_i^*(0)+\lambda_i^*(0)=0,$$

即

$$u_i(0)=-R_i^{-1}B_i^{\mathrm{T}}p_i^*(0),\tag{3-57}$$
$$z_i(0)=-S_i^{-1}(C_i^{\mathrm{T}}p_i^*(0)+\lambda_i^*(0)).$$

(2) 对于 $k=1,2,\cdots,K-1$,

$$\min_{x_i(k),u_i(k),z_i(k)}\Big\{H_i(x_i(k),u_i(k),z_i(k),k)-p_i^*(k-1)^{\mathrm{T}}x_i(k)\Big\},$$

其解为

$$x_i(k)=-Q_i(k)^{-1}\Big[A_i^{\mathrm{T}}p_i^*(k)-p_i^*(k-1)-\sum_{j=1}^{N}(\lambda_j^{*\mathrm{T}}L_{ji})^{\mathrm{T}}\Big],\tag{3-58}$$
$$u_i(k)=-R_i^{-1}B_i^{\mathrm{T}}p_i^*(k),$$
$$z_i(k)=-S_i^{-1}(C_i^{\mathrm{T}}p_i^*(k)+\lambda_i^*(k)).$$

(3) 对于 $k=K$,

$$\min_{x_i(K)}\Big\{\frac{1}{2}\|x_i(K)\|_{G_i}^2-p_i^{*\mathrm{T}}(K-1)x_i(K)\Big\},$$

其解为

$$x_i(K) = G_i^{-1}p_i^*(K-1). \tag{3-59}$$

图 3-4 所示的为离散系统三级递阶算法的结构图.

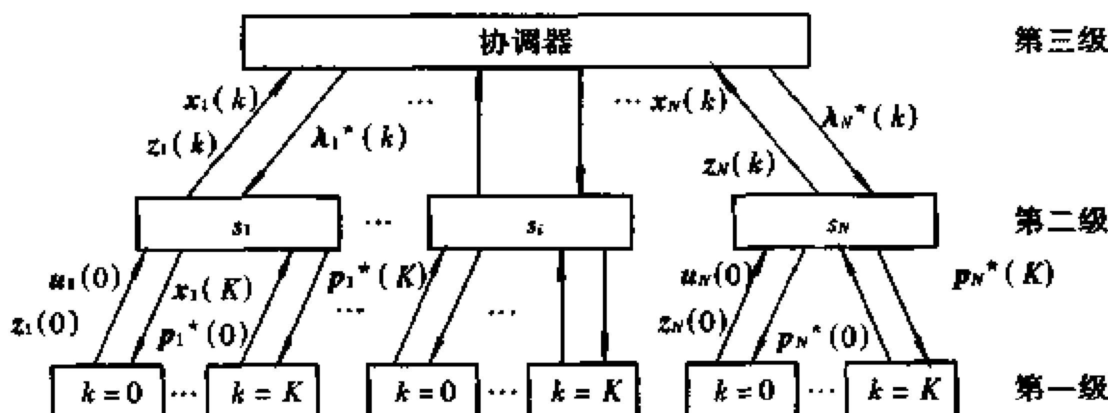


图 3-4

4　大系统分散控制

分散控制的主要特点是,将复杂的大系统按其分布划分成许多独立子系统,分别用独立的局部控制器来控制.每个控制器只观测系统的局部输出,且只控制系统的局部输入,共同完成大系统所要达到的目标.

分散控制系统分为分散随机控制系统和分散确定性控制系统.

由于分散控制系统具有非经典信息结构(见本篇第 1 章"大系统结构"),它使得分散控制系统最优决策复杂化,成为非线性的.即使对于最简单线性二次型高斯(G.F.Gauss)(LQG) 问题,分离定理也不再成立,反馈律也是非线性的.只有在特定的信息结构,例如一步时延共享信息结构或具有嵌套的信息结构下,分离定理才成立,并存在唯一线性最优解.

在分散确定性控制系统中,其研究的理论主要包括分散系统的状态估计、能控性、能观性、稳定性和分散控制系统的镇定和极点配置、分散最优控制等.20 世纪 70 年代初,戴维逊(Davision) 和王(Wang) 提出了分散固定模的概念,为系统地分析和设计分散确定性控制系统提供了重要基础.以此为依据,又发展了各种扰动和参数摄动的分散鲁棒控制和结构摄动的分散鲁棒控制的理论和方法.

本节就大系统分散控制的主要内容作简要介绍.

4.1　经典信息结构与非经典信息结构

分散控制的主要特征是,在每个时刻,各分散控制器(或称控制站)只能获得整个大系统的某一局部信息,并利用这种局部信息作出自己的控制决策;它只是整体控制作用中的某一局部控制作用,而且控制器的控制决策决定于其他控制器的

控制决策.这种在一个给定时刻,各控制器的动作取决于来自别的控制器信息的结构,称之为非经典信息结构.在数学上,可以描述如下:

考虑线性随机系统:

$$\dot{x}(t) = Ax(t) + \sum_{i=1}^{N} B_i u_i(t) + \xi(t), \tag{4-1}$$

$$y(t) = cx(t) + \sum_{i=1}^{N} D_i u_i(t) + \theta(t). \tag{4-2}$$

方程(4-1)式是受 N 个控制站控制的线性动态系统的状态方程.方程(4-2)式是输出方程.其中 $\theta(t),\xi(t)$ 假定为独立的高斯白色噪声过程;$y(t)$ 相当于系统的全局测量输出.这些对一个给定的控制器有用的测量的划分,是由问题的信息模式来确定的.对于分散控制问题,其信息模式定义为矩阵的集合:

$$H = (H_1, H_2, \cdots, H_n), \tag{4-3}$$

其中

$$z_i(t) = H_i(t)y(t) \qquad (i = 1,2,\cdots,N), \tag{4-4}$$

$z_i(t)$ 为分散控制器的信息.对于经典信息情况,定义 B 为对角方块阵(B_i),$i = 1, 2,\cdots,N$,D 为对角方块阵$[D_i]$,$i = 1,2,\cdots,N$.H 为对角方块阵$[H_i]$,$i = 1,2,\cdots,N$.当 $H = I$,即 H 为单位阵时,就产生了经典信息结构模式,这说明所有控制器都具有相同的信息.

如果考虑二次型性能指标

$$J = \lim_{T}\left\{ \frac{1}{T}\int_0^T \left[x^{\mathrm{T}}(t)Qx(t) + \sum_{i=1}^{N} u_i^{\mathrm{T}}(t)R_i u_i(t)\right] \mathrm{d}t \right\}, \tag{4-5}$$

其中 Q 为正半定矩阵,R 为对角方块正定矩阵,且如果控制限定为

$$u_i = \gamma_i(z_i^t),$$

其中

$$z_i^t = \{z_i(\tau), 0 \leqslant \tau \leqslant t\},$$

就产生了非经典信息结构,即第 i 个控制站的控制是它能得到的过去和现在信息的某个函数 γ_i.

非经典信息结构给分散控制带来的主要困难在于:即使在最简单的线性二次型高斯(LQG)问题中,著名的分离定理也不再成立,反馈律一般是非线性的.

4.2 分散确定性控制

在分散确定性控制中一个最常用也是最主要的概念是关于固定模的概念,它对于分散控制系统的稳定性、能控性、能观性的分析,以及分散镇定方法都具有重要作用,这是王和戴维逊于 1973 年提出来的.

4.2.1 问题的建立

考虑一个具有 N 个局部控制站的大规模线性时不变系统:

$$\dot{x}(t) = Ax(t) + \sum_{i=1}^{N} B_i u_i(t), \tag{4-6}$$

$$y_i(t) = C_i x(t) \quad (i = 1,2,\cdots,N). \tag{4-7}$$

其中 $x \in \mathbf{R}^n, u_i \in \mathbf{R}^{m_i}, y_i \in \mathbf{R}^{r_i}$ 分别表示系统状态向量、第 i 控制站的输入向量和输出向量.其原始系统的控制和输出阶次为 m 和 r,

$$m = \sum_{i=1}^{N} m_i, \quad r = \sum_{i=1}^{N} r_i, \tag{4-8}$$

A, B_i 和 C_i 是具有相应阶次的实常数矩阵.

分散镇定问题就是要找出 N 个具有动态补偿器的局部输出控制规律:

$$\begin{cases} \dot{z}_i(t) = S_i z_i(t) + R_i y_i(t), \\ u_i(t) = Q_i z_i(t) + K_i y_i(t) + v_i(t) \quad (i = 1,2,\cdots,N), \end{cases} \tag{4-9}$$

使得整个系统稳定.其中 $z_i(t) \in E^{\eta_i}$ 是第 i 个补偿器的输出向量;$u_i(t) \in \mathbf{R}^{m_i}$ 是第 i 个控制器外部输入向量;矩阵 S_i, R_i, Q_i, K_i,分别为 $\eta_i \times \eta_i, \eta_i \times r_i, m_i \times \eta_i, m_i \times r_i$ 阶矩阵.可以把(4-9)式写成更紧凑的形式:

$$\dot{z}(t) = Sz(t) + Ry(t), \tag{4-10}$$

$$u(t) = Qz(t) + Ky(t) + v(t), \tag{4-11}$$

其中

$$S \overset{\text{def}}{=\!=} \mathrm{diag}[S_1, S_2, \cdots, S_N],$$

$$Q \overset{\text{def}}{=\!=} \mathrm{diag}[Q_1, Q_2, \cdots, Q_N],$$

$$R \overset{\text{def}}{=\!=} \mathrm{diag}[R_1, R_2, \cdots, R_N],$$

$$K \overset{\text{def}}{=\!=} \mathrm{diag}[K_1, K_2, \cdots, K_N].$$

把反馈规律(4-10)式、(4-11)式用到系统(4-6)式和(4-7)式上,那么整个闭环系统就可以表示为

$$\begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} A + BKC & BQ \\ RC & S \end{bmatrix} \begin{bmatrix} x(t) \\ z(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} v(t), \tag{4-12}$$

其中

$$C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_N \end{bmatrix},$$

$$B = [B_1, B_2, \cdots, B_N].$$

如前面所指出的那样,问题是要求出控制规律(4-10)式、(4-11)式,使增广系统(4-12)式能渐近稳定.换句话说,用局部输出反馈,使分散系统的闭环极点处在复平面的左半平面中.

4.2.2 固定模与固定多项式

定义1 考虑由(4-6)式和(4-7)式所描述的系统(C,A,B)及整数$m_i,r_i,i=1,2,\cdots,N$.令$m\times r$的增益矩阵K代表下述对角方块矩阵集合中的矩阵：

$$\mathscr{K}=\left\{K\;\middle|\;K=\begin{pmatrix}K_1 & & & \\ & K_2 & & \\ & & \ddots & \\ & & & K_N\end{pmatrix}\right\},\tag{4-13}$$

其中$\dim(K_i)=m_i\times r_i,i=1,2,\cdots,N$.那么系统$(C,A,B)$对于$\mathscr{K}$的"固定多项式"就是对于所有$K\in\mathscr{K}$的多项式$|\lambda I-A-BKC|$集合的最大公共因子(g.c.d),并用下式表示：

$$\Phi(\lambda;C,A,B,K)=\underset{K\in\mathscr{K}}{\gcd}\{|\lambda I-A-BKC|\}.\tag{4-14}$$

定义2 对于系统(C,A,B)和由(4-13)式所给定的输出反馈的集合,$\mathscr{K}$的(C,A,B)的固定模的集合定义为矩阵$[A+BKC]$特征值所有可能集合的交集,即

$$\Lambda(C,A,B,\mathscr{K})=\bigcap_{K\in\mathscr{K}}\lambda(A+BKC),\tag{4-15}$$

其中$\lambda(\cdot)$表示$[A+BKC]$特征值的集合.注意到K可以取零矩阵,因此,固定模的集合包含于$\lambda(A)$中.根据定义1,固定模$\Lambda(\cdot)$即是(4-14)式中$\Phi(\cdot;\cdots)$固定多项式的根,即

$$\Lambda(C,A,B,K)=\{\lambda\mid\lambda\in S\text{和}\Phi(\lambda;C,A,B,\mathscr{K})=0\},\tag{4-16}$$

其中S表示整个S复平面上点的集合.

4.2.3 分散闭环系统稳定性的充分必要条件

定理1 对于(4-6)式和(4-7)式所描述的系统和(4-13)式那种类型的对角方块矩阵$\mathscr{K}$,当且仅当$(C,A,B,\mathscr{K})$固定模的集合包含在S复平面的左半开平面,即

$$\Lambda(C,A,B,\mathscr{K})\in S^-,\tag{4-17}$$

其中S^-表示S复平面的左半开平面,那么(4-9)式所表示的局部反馈规律才能使系统渐近稳定.

4.3 分散随机控制

4.3.1 离散时间随机大系统分散控制问题的建立

考虑一个离散时间随机大系统：

$$x(k+1)=Ax(k)+\sum_{i=1}^{N}B_iu_i(k)+\xi(k).\tag{4-18}$$

其中 $x(k)$ 是在阶段 k，$n\times 1$ 的状态向量；$u_i(k)$ 是第 i 个控制器在阶段 k，$m_i\times 1$ 的控制向量；$A, B_i, i=1,2,\cdots,N$，是已知常数矩阵；$\xi(k)$ 是零均值和已知方差的噪声，

$$E\{\xi(k)\xi^{\mathrm{T}}(k)\} = V(k)\delta(k,l), \tag{4-19}$$

其中 $\delta(\cdot,\cdot)$ 是克罗内克(Z.Kronecker)符号，$V(\cdot)$ 是方差，$E\{\cdot\}$ 代表期望值. 每一结点或通道的测量遵循下式：

$$y_i(k) = C_i x(k) + \eta_i(k). \tag{4-20}$$

其中 $y_i(k)$ 是在阶段 k，在 i 结点上一个 $r_i\times 1$ 的测量向量；C_i 是 $r_i\times n$ 矩阵；$\eta_i(k)$ 是零均值白色噪声. 假定各结点互不相关，已知方差

$$E\{\eta_i(k)\eta_i^{\mathrm{T}}(k)\} = W_i(k)\delta(k,l), \tag{4-21}$$

其中 $W(k)$ 为正定矩阵. 问题是：求 N 个分散控制器 $u_i(k), i=1,2,\cdots,N, k=1,2,\cdots,K$，使(4-18)式 ~ (4-21)式得到满足，并使以下特性指标最小：

$$J = E\left\{\frac{1}{2}\sum_{k=1}^{K}\left[x^{\mathrm{T}}(k)Q(k)x(k) + \sum_{i=1}^{N}u_i^{\mathrm{T}}(k)R_i(k)u_i(k)\right]\right\}, \tag{4-22}$$

其中 $R(k)$ 为正定矩阵.

4.3.2 分散解(离散形式)

分散随机控制的求解首先把状态估计 $\hat{x}(k)$ 分散解成两部分：决定于输入数据部分 $\hat{x}^{\mathrm{D}}(k)$ 和决定于控制部分 $x^{\mathrm{C}}(k)$，即

$$\hat{x}(k) = \hat{x}^{\mathrm{D}}(k) + x^{\mathrm{C}}(k). \tag{4-23}$$

其中 $x^{\mathrm{C}}(k)$ 由下式给定：

$$\begin{cases} x^{\mathrm{C}}(k+1) = Ax^{\mathrm{C}}(k) + \sum_{i=1}^{N}B_i u_i(k), \\ x^{\mathrm{C}}(1) = \bar{x}(1). \end{cases} \tag{4-24}$$

$\bar{x}^{\mathrm{D}}(1) = 0$，因此

$$\hat{x}^{\mathrm{D}}(k+1) = \bar{x}^{\mathrm{D}}(k+1) + \sum_{i=1}^{N}K_i(k+1)(\tilde{y}_i(k+1) - C_i\bar{x}^{\mathrm{D}}(k+1)), \tag{4-25a}$$

$$\bar{x}^{\mathrm{D}}(k+1) = A\hat{x}^{\mathrm{D}}(k), \tag{4-25b}$$

其中

$$\tilde{y}_i(k) \overset{\mathrm{def}}{=\!=} y_i(k) - C_i x^{\mathrm{C}}(k). \tag{4-26}$$

通过分解，可以证明：分散系统是由各站处理传感器数据的卡尔曼(R.E.Kalman)滤波器组成的. 令 $\hat{x}_i^{\mathrm{D}}(k) = E\{x^{\mathrm{D}}(k)/Y_i(k)\}$ 为在 i 站只用测量 $Y_i(k)$ 的局部估计，令 $P_i(k) = [(\hat{x}^{\mathrm{D}}(k) - x_i^{\mathrm{D}}(k)](\hat{x}^{\mathrm{D}}(k) - x_i^{\mathrm{D}}(k))^{\mathrm{T}}/Y_i(k)]$ 表示由 $Y_i(k)$ 引起的误差方差，利用附加的局部数据相关向量 $h_i(k)$，在给定所有数据情况下，状态估计

给定为

$$\hat{x}^{D}(k) = \sum_{i=1}^{N}[P(k)P_i^{-1}(k)\hat{x}_i^{D}(k) + h_i(k)], \tag{4-27}$$

其中

$$\hat{x}_i^{D}(k) = \bar{x}_i^{D}(k) + P_i(k)C_i^{T}W_i^{-1}(k)[\tilde{y}_i(k) - C_i\bar{x}_i^{D}(k)], \tag{4-28a}$$

$$\bar{x}_i^{D}(k) = Ax_i^{D}(k-1), \tag{4-28b}$$

$$P_i^{-1}(k) = M_i^{-1}(k) + C_i^{T}W_i^{-1}(k)C_i, \tag{4-29}$$

$$h_i(k+1) = F(k+1)h_i(k) + G_i(k+1)\bar{x}_i^{D}(k+1),$$

$$h_i(1) = 0, \tag{4-30}$$

$$F(k) = [I - \sum_{i=1}^{N}P(k)C_i^{T}W_i^{-1}(k)C_i]A = P(k)M^{-1}(k)A, \tag{4-31}$$

$$G_i(k+1) = P(k+1)M^{-1}(k+1)AP(k)P_i^{-1}(k)A^{-1}P(k+1)M_i^{-1}(k+1), \tag{4-32}$$

$$M(k+1) \xlongequal{\text{def}} E\left\{\frac{(\bar{x}(k+1) - x(k+1))(\bar{x}(k+1) - x(k+1))^{T}}{Y_1(k)\cdots Y_N(k)}\right\},$$

其中

$$Y_i(k) \xlongequal{=\text{def}} \{y_i(l), l = 1,2,\cdots,k\} \quad (i = 1,2,\cdots,N).$$

假定 A 为非奇异阵，当信号在结点 i 传送到结点 $l = 1,2,\cdots,i-1,i+1,\cdots,N$ 时，定义

$$\beta_i^{l}(k) = B_l^{T}S(k+1)[P(k)P_i^{-1}(k)\hat{x}_i^{D}(k) + h_i(k)], \tag{4-33}$$

它代表 $m_i \times 1$ 的向量. $S(k)$ 是时间向后传播，控制向量

$$u_i(k) = -(R_i + B_i^{T}S(k+1)B_i)^{-1}\left\{\sum_{l=1}^{N}\beta_l^{i}(k) + B_i^{T}S(k+1)x^{C}(k)\right\}. \tag{4-34}$$

4.3.3 连续时间随机大系统的分散控制

问题是寻求 N 个局部控制 $u_i(t), i = 1,2,\cdots,N$，使之满足随机差分方程

$$\mathrm{d}x = \left[Ax(t) + \sum_{i=1}^{N}B_iu_i(t)\right]\mathrm{d}t + \mathrm{d}\xi, \tag{4-35}$$

并使以下目标泛函极小：

$$J = E\left\{\frac{1}{2}\int_{t_0}^{t_f}\left[x^{T}(t)Q(t)x(t) + \sum_{i=1}^{N}u_i^{T}(t)R_i(t)u_i(t)\right]\mathrm{d}t\right\}. \tag{4-36}$$

局部输出测量为

$$\mathrm{d}y_i = C_i(t)x(t)\mathrm{d}t + \mathrm{d}\eta_i, \tag{4-37}$$

其中 ξ_i 和 $\eta_i, i = 1,\cdots,N$，是向量布朗(R. Brown)过程，具有零均值和方差：

$$E\{\mathrm{d}\xi\mathrm{d}\xi^{T}\} = V(t)\mathrm{d}t,$$

$$E\{d\boldsymbol{\eta}_i d\boldsymbol{\eta}_i^{\mathrm{T}}\} = \boldsymbol{W}_i(t)dt. \tag{4-38}$$

(4-36) 式中矩阵 $\boldsymbol{Q}(t)$ 和 $\boldsymbol{R}(t)$ 分别假定为对称半正定阵和正定阵. 测量历史定义为

$$\boldsymbol{Y}_i(t) \xlongequal{\text{def}} \{d\boldsymbol{y}_i(\tau); 0 \leqslant \tau \leqslant t\} \quad (i = 1,2,\cdots,N),$$

最优控制具有以下一般形式:

$$\boldsymbol{u}_i(t) = \boldsymbol{\varphi}_i(\boldsymbol{Y}_1(t), \boldsymbol{Y}_2(t), \cdots, \boldsymbol{Y}_N(t), t), \tag{4-39}$$

即 $\boldsymbol{u}_i$(控制器) 是从整个网络中所得信息的函数.

在分散控制中, 假定每站的数据都由它们自己的卡尔曼滤波器处理, 即

$$d\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(t) = \boldsymbol{A}\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(t)dt + \boldsymbol{P}_i(t)\boldsymbol{C}_i^{\mathrm{T}}(t)\boldsymbol{W}_i^{-1}(t)(d\widetilde{\boldsymbol{y}}_i(t) - \boldsymbol{C}_i(t)\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(t)dt), \tag{4-40}$$

其中

$$\widetilde{\boldsymbol{y}}_i(t) \xlongequal{\text{def}} \boldsymbol{y}_i(t) - \boldsymbol{C}_i(t)\boldsymbol{x}^{\mathrm{C}}(t). \tag{4-41}$$

局部误差方差 $\boldsymbol{P}_i(t)$ 只基于信息 $\boldsymbol{Y}_i(t)$, 而且传播为

$$\begin{aligned}\dot{\boldsymbol{P}}_i(t) = {} & \boldsymbol{A}(t)\boldsymbol{P}_i(t) + \boldsymbol{P}_i(t)\boldsymbol{A}^{\mathrm{T}}(t) + \boldsymbol{V}(t) - \\ & \boldsymbol{P}_i^{\mathrm{T}}(t)\boldsymbol{C}_i^{\mathrm{T}}(t)\boldsymbol{W}_i^{-1}(t)\boldsymbol{C}_i(t)\boldsymbol{P}_i(t)(\text{对于给定的 } \boldsymbol{P}_i(0)).\end{aligned} \tag{4-42}$$

一旦所有数据具备, 状态估计为

$$\widehat{\boldsymbol{x}}^{\mathrm{D}}(t) = \sum_{i=1}^{N}\{\boldsymbol{P}(t)\boldsymbol{P}_i^{-1}(t)\widehat{\boldsymbol{x}}_i^{\mathrm{D}} + \boldsymbol{h}_i(t)\}, \tag{4-43}$$

其中 $\boldsymbol{h}_i(t)$ 的动态方程由下式给定:

$$\begin{aligned}\dot{\boldsymbol{h}}_i(t) = {} & [\boldsymbol{A}(t) - \sum_{i=1}^{N}\boldsymbol{P}(t)\boldsymbol{C}_i^{\mathrm{T}}(t)\boldsymbol{W}_i^{-1}(t)\boldsymbol{C}_i(t)]\boldsymbol{h}_i(t) + \\ & [\boldsymbol{P}(t)\boldsymbol{P}_i^{-1}(t) - \boldsymbol{I}]\boldsymbol{V}(t)\boldsymbol{P}_i^{-1}(t)\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(t).\end{aligned} \tag{4-44}$$

由于 $\widehat{\boldsymbol{x}}^{\mathrm{D}}(0)$ 和 $\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(0)$ 是零向量, 因此

$$\boldsymbol{h}_i(0) = 0,$$

从 $\boldsymbol{h}_i(t)$ 连续时间公式可以看出, $\boldsymbol{h}_i(t)$ 对过程噪声方差的依赖性. 如果对于 $0 \leqslant t \leqslant t_f$, $V(t) = 0$, 则 $\boldsymbol{h}_i(t) = 0$, (4-44) 式的附加项有稳定作用, 因为它是负半定的.

与离散时间系统相似, 连续时间的控制器为

$$\boldsymbol{u}_i(t) = -\boldsymbol{R}_i^{-1}(t)\boldsymbol{B}_i^{\mathrm{T}}(t)\boldsymbol{S}(t)\{\sum_{i=1}^{N}[\boldsymbol{P}(t)\boldsymbol{P}_i^{-1}(t)\widehat{\boldsymbol{x}}_j^{\mathrm{D}}(t) + \boldsymbol{h}_j(t)] + \boldsymbol{x}^{\mathrm{C}}(t)\}, \tag{4-45}$$

因此, 在每个 $(N-1)$ 站, 具有以下形式的连续时间信息:

$$\boldsymbol{\beta}_i^j(t) = \boldsymbol{B}_j^{\mathrm{T}}(t)\boldsymbol{S}(t)[\boldsymbol{P}(t)\boldsymbol{P}_i^{-1}(t)\widehat{\boldsymbol{x}}_i^{\mathrm{D}}(t) + \boldsymbol{h}_i^j(t)], \tag{4-46}$$

将从 $j = 1,2,\cdots,l-1,l+1,\cdots,N$ 的其他站发送到每一站.

4.4 队 论

在分散决策的情况下, 有许多决策者. 每个决策者在不同时刻控制不同的决策

变量.队是一个组织,在这个组织里,对所有的成员,有一个共同的单一目标或价值函数.何毓琦指出,队论的主要要素为

(1) 对每一个决策者来说,都拥有某些不确定性的、不同的、相关的信息;

(2) 为了获得性能指标,所有决策者的作用需要协调.

如果缺掉其中的一个或一个以上的要素,则问题简化成为去耦或变得平凡了.

须要强调的是,允许决策者预先交换信息并达成协议,也即他们对所采取的协调动作可取得一致意见.该协调动作是他们各自取得信息的函数.这与传统的非协调博弈是不同的,在那里不可能有强制的预定协议.

4.4.1　队决策的模型

由何毓琦所提出的模型中,有五个基本元素:

(1) 具有给定分布 $p(\xi)$ 的随机向量 $\xi,\xi=(\xi_1,\xi_2,\cdots,\xi_m)$.这个随机向量代表了所有与问题有关的不确定性.例如测量噪声、随机干扰、不确定的随机干扰等.ξ 常被当做"自然状态"或"自然决策".

(2) 一个观测集合 $z=\{z_1,z_2,\cdots,z_n\}$,它是 ξ 的给定函数,即

$$z_i=\eta_i(\xi_1,\xi_2,\cdots,\xi_m)\quad(i=1,2,\cdots,n).\tag{4-47}$$

z_i 通常是一个向量,并看做是第 i 个决策者(DM)所具有的信息或观测;集合 $\{\eta_i \mid i=1,2,\cdots,n\}$ 称为问题的信息结构.

(3) 每一决策者有一个决策变量集合 $\{u_1,u_2,\cdots,u_n\}\equiv u$.为不失一般性,可以假定 u_i 是一个标量.如果 u_i 是向量,它可以被分解成多于一个以上的决策者,并假定他们获得相同的观测 z_i.

变量 ξ,z,u 都假定在适当的空间 Θ,Z,U 中取值.

(4) 第 i 个决策者(DM)的策略(决策规律或控制规律)是一个映射:

$$\gamma_i:z_i\to U_i,$$

它代表在观测的基础上采取什么样决策的一种列联计划,因而可写作

$$u_i=\gamma_i(z_i).\tag{4-48}$$

(5) 问题的价值函数是一个映射:$L_i\quad\Theta\times U\to R$,即

$$\text{价值}=L(u_1,u_2,\cdots,u_n,\xi_1,\xi_2,\cdots,\xi_m).\tag{4-49}$$

注意,对于一个给定的策略集合 $\gamma_i,i=1,2,\cdots,n$,L 是一个 ξ 的确定函数,即

$$L(u,\xi)=L(\cdots,u_i=\gamma_i(\eta_i(\xi)),\xi_i,\cdots),$$

因而 L 对 $p(\xi)$ 的期望是充分确定的.

故可以把队决策问题叙述为

$$\begin{cases}\text{求 }\gamma_i\in\Gamma_i,\text{使}\\ J=E_i[L(u=\gamma(\eta(\xi),\xi)]\text{最小},\end{cases}\tag{4-50}$$

或

$$\min_{\gamma\in\Gamma}J(\gamma).$$

方程(4-50)式可看做是队问题的"策略"形式.从概念上,注意到方程(4-50)式是一个确定性的最优化问题.显然,一般来说这是一个难题.

4.4.2 部分嵌套信息站

在一般情况下,队决策的最优解是非线性的,而且是很难求得的.只有在特殊的信息结构下,才有线性最优解.如果有控制 $u_1,u_2,\cdots,u_N,u_{i+1}$ 知道 u_i 所知道的信息,则这种信息结构称为**部分嵌套信息结构**.具有部分嵌套信息结构的队决策,具有线性最优解.

5 大系统稳定性理论

判断、分析大系统能否正常稳定运行的理论和方法是大系统稳定性研究的主要内容.研究大系统稳定性问题,基本上有三个步骤:

(1) 把给定的大系统分解成若干小规模的子系统;

(2) 利用经典的稳定性理论和方法分析各子系统的稳定性;

(3) 把所得的结果合起来,并将各子系统之间的相互关联作为约束条件,推导整个系统稳定性的判据.

根据稳定性定义的不同,分析大系统稳定性的方法主要有两种:

(1) **李雅普诺夫函数法** 即对于各子系统,假设一个李雅普诺夫(Lyapunov)函数,然后利用向量李雅普诺夫函数理论,或者把各系统的李雅普诺夫函数加权求和,构造标量李雅普诺夫函数来检验整个系统的稳定性.

(2) **输入输出法** 即将各子系统在泛函空间上用一个数学表达式或算子来描述,然后用泛函分析方法来进行大系统稳定性分析.

在互连的大系统稳定性分析中,主要的研究的问题是:互作用的幅度和强度究竟可以达到多大,才不影响整个系统的稳定性.

5.1 李雅普诺夫方法

5.1.1 定义和问题描述

定义1 一个状态 x 和时间 t 的标量实值函数 $v(x,t)$ 被认为是"正定"的,如果有一个非递减实值函数 $\hat{v}(x,t)$,使得对于所有 $x\neq 0,\hat{v}(0,0)=0,0<\hat{v}(x,t)\leqslant v(x,t)$.如果 $\hat{v}(\infty,t)=\infty$,则函数 $v(x,t)$ 进一步称为径向无界.符号 $\|\cdot\|$ 表示欧几里德(Euclidean)范数,定义为 $\|x\|=(x_1^2+x_2^2+\cdots+x_n^2)^{1/2}$.

定义2 一个正定函数 $v(x,t)$ 称为"递减的",如果存在一个非递减函数 $v(\|x\|)$,使得 $\bar{v}(0)=0,v(x,t)\leqslant\bar{v}(\|x\|)$.正定、递减和径向无界函数被表示为 pdu.

定义3 一个 $n\times n$ 矩阵 A 称为"M"或"Metgler"矩阵,如果以下等价条件满足:

1° A 的所有主子式为正;

2° A 的所有前主子项为正；

3° 对于元素都为正的向量 x(或 y)，Ax(或 $A^{\mathrm{T}}x$) 的元素都为正；

4° A^{-1} 存在，其所有元素为非负；

5° $\mathrm{Re}\{\lambda_i(A)\} > 0$，对于 $i = 1,2,\cdots,n$；

6° 对角矩阵 $B = \mathrm{diag}[b_1, b_2, \cdots, b_n]$，$b_i > 0$，存在，使得 $BA + B^{\mathrm{T}}B$ 为正定矩阵，这个条件有时称为李雅普诺夫型.

定义 4 一个实平方矩阵 Ω 称为"M"矩阵，如果有一个对角矩阵

$\tilde{D} = \mathrm{diag}[\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_n]$，$\tilde{d}_i > 0$，使得 $\tilde{D} - \Omega^{\mathrm{T}}D\Omega$ 是正定矩阵.

现在考虑非强迫的大系统：

$$\mathrm{L}_{\mathrm{ss}}: \dot{x} = f(x, t). \tag{5-1}$$

它由 N 个子系统组成：

$$\mathrm{C}_{\mathrm{ss}}: \dot{x}_i = f_i(x_i, t) + g_i(x, t) \quad (i = 1,2,\cdots,N). \tag{5-2}$$

其中 x_i，$f_i(\cdot)$ 和 $g_i(\cdot)$ 满足 $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, \cdots, x_N^{\mathrm{T}})$，$f = \{f_1^{\mathrm{T}}(\cdot) + g_1^{\mathrm{T}}(\cdot), \cdots, f_N^{\mathrm{T}}(\cdot) + g_N^{\mathrm{T}}(\cdot)\}$；$\mathrm{L}_{\mathrm{ss}}$ 和 C_{ss} 分别表示原始大系统和复合子系统. 进一步假定原点 $x_{\mathrm{e}} = 0$ 是一个平衡状态，即对于 $i = 1,2,\cdots,N$，

$$f(0,t) = 0, \quad f_i(0,t) = 0, \quad g_i(0,t) = 0. \tag{5-3}$$

复合子系统的集合可重写为

$$\dot{x}_i = f_i(x_i, t) + u_i, \tag{5-4}$$

其中 $u_i = g_i(x, t)$，实际上代表了交互或对第 i 子系统的交互连接输人. 为了便于讨论，当第 i 个子系统完全去耦时，称

$$\mathrm{L}_{\mathrm{ss}}: \dot{x}_i = f_i(x_i, t) \tag{5-5}$$

为孤立系统. 而且现在的讨论只限于相对于平衡点 $x_{\mathrm{e}} = 0$ 或 $x_{i\mathrm{e}} = 0$，$i = 1,2,\cdots,N$. 下面定义一致稳定、一致渐近稳定和李雅普诺夫意义下一致渐近稳定.

定义 5 对于一个正定、递减系统状态 x 和时间 t 的函数 $v(x,t)$，(5-1) 式的平衡点被称为"一致稳定"，如果对于所有 x 和 t，$-\dot{v}(x,t)|_{\mathrm{L_{ss}}} \geqslant 0$. 注意符号 $k(\cdot)|_{\mathrm{L_{ss}}}$ 意味着变量 $k(\cdot)$ 是沿大系统(5-1) 式的轨迹进行计算的.

定义 6 按定义 5，平衡点 $x_{\mathrm{e}} = 0$ 是"一致渐近稳定"的，如果 $-\dot{v}(x,t)|_{\mathrm{L_{ss}}} > 0$ 和"基本上一致渐近连接稳定"，且 $v(x,t)$ 是径向无界(定义 1) 的. 定义 5 和定义 6 的函数 $v(x,t)$ 称为**李雅普诺夫函数**.

大系统稳定问题表达为：对于定义 1 的大系统，它可以分解为具有孤立子系统 L_{ss}((5-5) 式) 的 N 个复合子系统 L_{ss}((5-4) 式) 和互连接项 u_i(定义 4)，那么，在什么条件下原始复合系统是一致渐近连接稳定的呢？

5.1.2 大系统李雅普诺夫稳定条件

定理 1 由组合子系统(5-2) 式所代表的大系统是基本一致渐近稳定的，如果以下条件成立：

1° 对于每一个 I_{∞},存在一个 pdu(定义 2),使

$$\dot{v}_i(x_i,t)\Big|_{I_\infty} \leqslant -a_i\{W_i(x_i)\}^2, \tag{5-6}$$

$$\left|\frac{\partial v_i(\cdot)}{\partial x_i}\right| \leqslant W_i(x_i), \tag{5-7}$$

其中 $a_i, i=1,2,\cdots,N$,是正常数,$W_i(x_i)$ 是正定函数.

2° 互连接项由下式界定:

$$|g_i(x,t)| \leqslant \sum_{j=1}^{N} b_{ij}W_j(x_j), \tag{5-8}$$

其中 b_{ij} 是非负常数.

3° 由

$$e_{ij} = a_i - b_{ii}, e_{ij} = -b_{ij} \quad (i \neq j), \tag{5-9}$$

给定的 $N \times N$ 矩阵 $E=[e_{ij}]$ 是一个 M 矩阵,即 E 前主子项都为正,

$$D_i \xlongequal{\text{def}} \det\begin{bmatrix} e_{11} & \cdots & e_{1i} \\ \cdots & \cdots & \cdots \\ e_{i1} & \cdots & e_{ii} \end{bmatrix} > 0 \quad (i=1,2,\cdots,N). \tag{5-10}$$

上述条件中要求所有独立子系统都是基本一致渐近稳定的.这个假定颇为严格.下面的判据要求的条件较弱.

由组合子系统(5-2)式所代表的大系统是基本一致渐近稳定的,如果以下条件成立:

1° 对于每一个 I_{∞},存在一个 pdu 函数 $v_i(x_i,t)$,使得

$$\dot{v}_i(x_i,t)\Big|_{I_\infty} \leqslant -a_i\{W_i(x_i)\}^2 - z_i(x_i). \tag{5-11}$$

其中 $a_i, i=1,2,\cdots,N$,是正常数;$W_i(x_i)$ 和 $z_i(x_i)$ 分别为正半定和正定函数.

2° 互连接项满足

$$\left(\frac{\partial v_i(x_i,t)}{\partial x_i}\right)^{\mathrm{T}} g_i(x,t) \leqslant W_i(x_i)\sum_{j=1}^{N} b_{ij}W_j(x_j). \tag{5-12}$$

3° 由(5-9)式所定义的 $N \times N$ 矩阵 E 是一个 M 阵.

5.2 输入输出稳定方法

5.2.1 问题描述

输入输出稳定是建立在函数空间基础上的.令 $U^{(\mu)}$ 表示 $\mathbf{R}^{\mu}$ 时间值函数 $u(t)$ 的赋值空间,则"扩充赋值空间"$U_e^{(\mu)}$ 定义为

$$U_e^{(\mu)} = \{u \mid u_\tau \in U^{(\mu)}\ \forall \tau \in \mathbf{R}\},$$

其中 u 是"截函数",定义为

$$u_\tau(t) = \begin{cases} u(t) & (t < \tau), \\ 0 & (\text{其他}). \end{cases}$$

定义算子 G: $U^{(\mu)} \to U^{(\mu)}$,它被称为"输入输出稳定"或简单地叫 IO 稳定,如果存在2个非负常数 α 和 β,使得

$$\|(\mathrm{G}u)_\tau\| \leqslant \alpha\|u_\tau\| + \beta \quad (\text{对于所有 } \tau \in \mathbf{R}), \tag{5-13}$$

其中 $\|\cdot\|$ 是空间 $U^{(\mu)}$ 中的范数.

算子 G 的增益定义为

$$\mathrm{gain}\mathrm{G} = \max_{u \in U_e^{(\mu)}} \left\{ \frac{\|(\mathrm{G}u)_\tau\|}{\|u_\tau\|} \right\}. \tag{5-14}$$

令 D,F,G 和 H 是从 $U^{(\mu)}$ 到它本身的算子,使 G0 = 0,H0 = 0,现在考虑 IO 闭环系统:

$$e = \mathrm{D}u + \tilde{y},\ y = \mathrm{G}e,$$
$$\tilde{e} = \mathrm{F}u + y,\ \tilde{y} = \mathrm{H}\tilde{e}, \tag{5-15}$$

其中 u 是系统的输入.令 E,$\tilde{\mathrm{E}}$,Y 和 $\tilde{\mathrm{Y}}$ 分别代表 $u \in U^{(\mu)}$ 映射到解答 $e, \tilde{e}, y$ 和 $\tilde{y}$ 的算子,那么 IO 系统(5-15)式被认为是稳定的,如果算子 E,$\tilde{\mathrm{E}}$,Y 和 $\tilde{\mathrm{Y}}$ 是 IO 稳定的.因为绝大多数要处理的大系统要分解成 N 个子系统,所以,把 $U^{(\mu)}$ 当作 $U^{(\mu_i)}$ 的乘积是有用的,$i = 1,2,\cdots,N$,使 $\mu = \mu_1 + \mu_2 + \cdots + \mu_N$,且

$$\|u\| = \left(\sum_{i=1}^{N}\|u_i\|^2\right)^{1/2} \quad (u = (u_1^{\mathrm{T}}, u_2^{\mathrm{T}}, \cdots, u_N^{\mathrm{T}})^{\mathrm{T}}),$$
$$u_i \in U^{(\mu_i)}. \tag{5-16}$$

令 IO 系统(5-15)式由 N 个去耦子系统表示:

$$e_i = \mathrm{D}_i u + \tilde{y}_i,\ y_i = \mathrm{G}_i e_i,$$
$$\tilde{e}_i = \mathrm{F}_i u + y_i,\ \tilde{y} = \mathrm{H}_i\tilde{e}_i + K_i\tilde{e} \quad (i = 1,2,\cdots,N). \tag{5-17}$$

其中 $e = (e_1^{\mathrm{T}}, e_2^{\mathrm{T}}, \cdots, e_N^{\mathrm{T}})^{\mathrm{T}}, y = (y_1^{\mathrm{T}}, y_2^{\mathrm{T}}, \cdots, y_N^{\mathrm{T}})^{\mathrm{T}}$,等等;$e_i, \tilde{e}_i, y_i$ 和 $\tilde{y}_i$ 是 $U_e^{(\mu_i)}$ 的成员;G_i 和 H_i 是从 $U_e^{(\mu)}$ 到它本身的算子;$\mathrm{D}_i, \mathrm{F}_i$ 和 K_i 是 $U_e^{(\mu)}$ 到 $U_e^{(\mu_i)}$ 的算子.

大系统输入输出稳定问题就是研究在什么条件下,由(5-17)式表示的分解大系统是输入输出稳定的.

5.2.2 IO 系统稳定的必要条件

定理 2 (小增益定理) 输入输出系统(5-15)式是输入输出稳定的,如果

$$(\mathrm{gain\ G}) \cdot (\mathrm{gain\ H}) < 1, \tag{5-18}$$

其中 gain G 定义于(5-14)式.

定理 3 分解成(5-17)式的大系统(5-15)式是输入输出稳定的,如果以下条件满足:

1° 增益 $\mathrm{G}_i = \alpha_i < \infty, \quad (i = 1,2,\cdots,N)$. (5-19)

2° 对于一组 $2N$ 个非负常数 $\boldsymbol{\beta}_{ij}$ 的集合,以下范数条件成立:

$$\|(\mathrm{H}_i\tilde{e}_i + \mathrm{K}_i\tilde{e}_i)_\tau\| \leqslant \sum_{j=1}^{N}\beta_{ij}\|(\tilde{e}_i)_\tau\| \quad (\tilde{e} \in \|U_e^{(\mu)}\|, \tau \in \mathbf{R}). \tag{5-20}$$

3° $N \times N$ 的矩阵 $\boldsymbol{B} = [b_{ij}]$ 是一个 M 矩阵：

$$b_{ij} = 1 - \alpha_{ii}\beta_{ii}, b_{ij} = -\alpha_{ii}\beta_{ij} \quad (i \neq j).$$

5.3　复合系统的能控性与能观性

不失一般性，只考虑由 2 个子系统构成的复合大系统.

5.3.1　开环复合系统能控和能观的充要条件

考虑系统

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_1 & G \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u, \tag{5-21}$$

$$y = (C_1 \quad C_2)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + Du. \tag{5-22}$$

其中 x_1, x_2, u 和 y 分别为 n_1, n_2, m 和 r 维的子状态、控制和输出向量；子系统矩阵 A_1 和 A_2 分别为 $n_1 \times n_1$ 和 $n_2 \times n_2$；G 是 $n_1 \times n_2$ 的互连矩阵；C_1 和 C_2 是 $r \times n_1$ 和 $r \times n_2$ 输出矩阵；D 是 $r \times m$ 矩阵. 令 A_1 和 A_2 的特征值为

$$\lambda\{A_1\} = \lambda_i^1 \quad (i = 1,2,\cdots,n_1),$$

$$\lambda\{A_2\} = \lambda_i^2 \quad (i = 1,2,\cdots,n_2). \tag{5-23}$$

定理 4　(5-21) 式、(5-22) 式的能控的充要条件为

1° $\mathrm{rank}P_2(\lambda_i^2) = \mathrm{rank}[A_2 - \lambda_i^2 I, B_2] = n_2 \quad (i = 1,2,\cdots,n_2)$, (5-24)

2° $$\mathrm{rank}P_1(\lambda_i^1) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^1 I & G & B_1 \\ 0 & A_2 - \lambda_i^1 & B_2 \end{bmatrix} = n_1 + n_2 \quad (i = 1,2,\cdots,n_1). \tag{5-25}$$

进而，系统(5-21) 式、(5-22) 式能观的充要条件为

1° $$\mathrm{rank}Q_1(\lambda_i^1) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^1 I \\ C_1 \end{bmatrix} = n_1 \quad (i = 1,2,\cdots,n_1), \tag{5-26}$$

2° $$\mathrm{rank}Q_2(\lambda_i^2) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^2 I & G \\ 0 & A_2 - \lambda_i^2 I \\ C_1 & C_2 \end{bmatrix} = n_1 + n_2 \quad (i = 1,2,\cdots,n_2). \tag{5-27}$$

5.3.2　闭环复合系统能控和能观的充要条件

考虑闭环复合系统

$$\begin{cases} \dot{x}_1 = A_1 x_1 + B_1 e, & e = u - z, \\ \dot{x}_2 = A_2 x_2 + B_2 y, & z = C_2 x_2, \\ y = C_1 x_1 + D_1 e, \end{cases} \tag{5-28}$$

其中 e 是系统误差的测量，y 是总的输出，x_1 和 x_2 是子系统的状态，u 是控制．为检验此系统的能控性和能观性，可以消去(5-28)式中的 e 和 z，建立类似于(5-22)式的方程，即

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_1 & -B_1 C_2 \\ B_2 C_1 & A_2 - B_2 D_1 C_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 D_1 \end{bmatrix} u,$$

$$y = (C_1 \vdots -D_1 C_2) \begin{bmatrix} x_1 \\ x_1 \end{bmatrix} + D_1 u. \tag{5-29}$$

定理 5　闭环复合系统(5-29)式能控的充要条件为

$$1^\circ \ \mathrm{rank} P_1(\lambda_i^1) = \mathrm{rank}[A_1 - \lambda_i^1 I, B_1] = n_1 \quad (i = 1,2,\cdots,n_1), \tag{5-30}$$

$$2^\circ \ \mathrm{rank} P_2(\lambda_i^2) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^2 I & 0 & B_1 \\ B_2 C_1 & A_2 - \lambda_i^2 I & B_2 D_1 \end{bmatrix} = n_1 + n_2 \quad (i = 1,2,\cdots,n_2). \tag{5-31}$$

进而，系统(5-29)式能观的充要条件为

$$1^\circ \ \mathrm{rank} Q_1(\lambda_i^1) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^1 I \\ C_1 \end{bmatrix} = n_1 \quad (i = 1,2,\cdots,n_1), \tag{5-32}$$

$$2^\circ \ \mathrm{rank} Q_2(\lambda_i^2) = \mathrm{rank}\begin{bmatrix} A_1 - \lambda_i^2 I & -B_1 C_2 \\ 0 & A_2 - \lambda_i^2 I \\ C_1 & -D_1 C_2 \end{bmatrix} = n_1 + n_2 \quad (i = 1,2,\cdots,n_2). \tag{5-33}$$

参 考 文 献

1　李人厚，邵福庆．大系统递阶与分散控制．西安：西安交通大学出版社，1986．

2　Jamshidi M. Large-scale systems: modeling and control. New York: Elsevier Science Publishing Co Inc, 1983.

3　(英)辛格 M G. 分散控制．李人厚，胡保生译．北京：国防工业出版社，1985．

·经济数学卷·

第17篇

对策论

编　者　王建华
审校者　马振华

目　录

引　　言

对策论是研究两个或两个以上的参加者在竞争性或对抗性的局势下如何采取行动,作出有利于己方的决策的数学理论.这里的两个或多个参加者(称为局中人)相互之间有利害冲突,每个参加者只能控制部分局势,有时也不排除某些参加者之间有所合作,最终是各自作出抉择(选择一个策略),以得到一个对己方最为有利的结局.

对策论成为数学的一个分支,始于1944年.该年,冯·诺伊曼(J. von Neumann)和摩根斯顿(O. Morgenstern)出版了奠基性的经典著作《对策论与经济行为》.在该书中,第一次给对策以明确的数学描述,并且讨论了对策论在经济学中的一些应用.

从1944年到现在,已过去了50余年,对策论在理论和应用等各方面都有了不少发展.在理论方面,从最初的二人对策发展到 n 人对策,从离散对策发展到连续对策,从非合作对策发展到合作对策.在应用方面,从最初的经济学领域扩展到军事、政治、心理学、社会学等领域.在对策的计算方面,也取得了一些较优秀的成就.

1　矩阵对策

1.1　矩阵对策

1.1.1　矩阵对策基本概念

矩阵对策是整个对策论的研究基础,其中零和二人对策是最基本的对策.在二人对策中,一个局中人的赢得等于另一个局中人的失去的二人对策称为零和二人对策.下面举例说明.

1.配钱币游戏

二个参加者,称为局中人1和局中人2,各拿出一枚钱币.在不让对方看见的情况下,将钱币出示给对方.如果两个钱币都呈正面或都呈反面,则局中人1得1分,局中人2得 -1分,或者说,局中人2输给局中人1一个单位.如果两个钱币呈一正一反,则局中人1输给局中人2一个单位.可以用一个方阵来表示这些结果:

		局中人2	
		1(正)	2(反)
局中人1	1(正)	1	-1
	2(反)	-1	1

这种情况下,局中人1和局中人2各有两个策略.每个局中人的第1个策略表示选择出示钱币的正面,第2个策略表示选择钱币的反面.

上面的表格称为对策的**支付矩阵**,它是两个局中人的策略的函数.局中人1的策略用支付矩阵的行来表示,局中人2的策略用支付矩阵的列来表示.每个局中人选定一个策略后,支付矩阵中有一个对应的元素,它代表局中人2应当付给局中人1的支付值.例如,若局中人1出反面(策略2),局中人2也出反面(策略2),则在上表中第2行第2列处的元素1就是局中人2应该付给局中人1的数目.这时,局中人1得到支付1,即局中人1赢进1个单位,局中人2输掉1个单位.正的支付值表示局中人1得到的支付是正的值.反之,负的支付值表示局中人1从局中人2处得到的是负的值,这就是说,局中人1失去若干个单位,这若干个单位被局中人2赢得.

这种游戏就是一个对策.所谓**对策**,就是一组规则,它描述整个游戏(或竞赛、或竞争、或斗争)自始至终所应遵循的各项规定,包括局中人、策略、选定策略后的支付等.

2."拳头、剪刀、布"游戏

每个小孩都玩过这个游戏.拳头胜剪刀、剪刀胜布、布胜拳头.这里也是两个局中人:局中人1和局中人2.每个局中人有三个策略:策略1代表出拳头,策略2代表出剪刀,策略3代表出布.假设胜者得1分,负者得-1分,则支付矩阵为

		局中人2		
		1	2	3
局中人1	1	0	1	-1
	2	-1	0	1
	3	1	-1	0

表中的元素是局中人1应得的支付值.

定义1 设局中人1有 m 个策略,$i=1,2,\cdots,m$;局中人2有 n 个策略,$j=1,2,\cdots,n$.设局中人1,局中人2分别选择策略 i 和 j 时,局中人1从局中人2处得到的支付是 a_{ij},则支付矩阵为

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & & \cdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}. \tag{1-1}$$

对策由上列矩阵完全确定.这种对策叫做 $m \times n$ **矩阵对策**. $\{1,2,\cdots,m\}$ 和 $\{1,2,\cdots,n\}$ 分别是局中人 1 和局中人 2 的**策略集**.

a_{ij} 是局中人 1 得到的支付,局中人 2 得到的支付则为 $-a_{ij}$.由于总共只有两个局中人,而且在任何情况下双方得到的支付之和为零,即 $a_{ij} + (-a_{ij}) = 0$,这种对策称为**零和二人对策**.

1.1.2 最优策略与鞍点

在以(1-1)式的 $A = [a_{ij}]$ 为支付矩阵的零和二人对策中,局中人 1 希望支付 a_{ij} 越大越好,而局中人 2 则刚好相反,他希望 $-a_{ij}$ 越大越好,即 a_{ij} 越小越好.但在支付矩阵$[a_{ij}]$中,每个局中人只能控制两个变量 i 和 j 中的一个:局中人 1 只能控制变量 i,局中人 2 只能控制变量 j.

如果局中人 1 选定一个策略 i,则他至少可以得到支付 $\min\limits_{1 \le j \le n} a_{ij}$.这就是支付矩阵第 i 行的最小元素.由于局中人 1 希望支付值越大越好,他可以选择 i 使上式为最大,即他可以选择 i 使支付不小于

$$\max_{1 \le i \le m} \min_{1 \le j \le n} a_{ij}. \tag{1-2}$$

同理,如果局中人 2 选定一个策略 j,则局中人 1 得到的支付不会超过 $\max\limits_{1 \le i \le m} a_{ij}$.这就是支付矩阵第 j 列的最大元素.由于局中人 2 希望支付值越小越好,他可以选择 j 使上式为最小,即他可以选择 j 使局中人 1 得到的支付不大于

$$\min_{1 \le j \le n} \max_{1 \le i \le m} a_{ij}. \tag{1-3}$$

(1-2) 式是支付矩阵各行最小值中的最大者,(1-3) 式是支付矩阵各列最大值中的最小者.容易看出,对于 1.1.1 小节的例子 1. 和 2. 的对策,都有

$$-1 = \max_{1 \le i \le 2} \min_{1 \le j \le 2} a_{ij} < \min_{1 \le j \le 2} \max_{1 \le i \le 2} a_{ij} = 1.$$

而对于以(例如)

$$A = \begin{bmatrix} 0 & 1 & 4 \\ -1 & 2 & 7 \\ -4 & 1 & 8 \\ -9 & -2 & 7 \end{bmatrix}$$

为支付矩阵的 4×3 矩阵对策,则有

$$\max_{1 \le i \le 4} \min_{1 \le j \le 3} a_{ij} = 0 = \min_{1 \le j \le 3} \max_{1 \le i \le 4} a_{ij}.$$

在一般情况,有如下定理.

定理 1 对于任意 $m \times n$ 矩阵对策 $A = [a_{ij}]$,必有

$$\max_{1 \le i \le m} \min_{1 \le j \le n} a_{ij} \leqslant \min_{1 \le j \le n} \max_{1 \le i \le m} a_{ij}. \tag{1-4}$$

当上式中等号成立时,即当

$$\max_{1\leqslant i\leqslant m}\min_{1\leqslant j\leqslant n} a_{ij} = v = \min_{1\leqslant j\leqslant n}\max_{1\leqslant i\leqslant m} a_{ij} \tag{1-5}$$

时,这个值 v 称为对策的值.此时,必有一个 $i = i^*$ 和一个 $j = j^*$,使

$$a_{ij^*} \leqslant a_{i^*j^*} = v \leqslant a_{i^*j} \tag{1-6}$$

对于一切 i 和一切 j 成立.

i^* 和 j^* 分别称为局中人1和局中人2的**最优策略**.(i^*, j^*) 是对策的一个**鞍点**,或称为对策的一个解.

从(1-6)式可以看出,对策在鞍点 (i^*, j^*) 处的值 v,既是支付矩阵中它所在的行中的最小元素,又是它所在的列中的最大元素.因此,对策如果有鞍点,很容易直接求出来.

鞍点有下列性质.

定理2　如果 (i^*, j^*) 和 (i^0, j^0) 都是 $m \times n$ 矩阵对策 $A = [a_{ij}]$ 的鞍点,则 (i^*, j^0) 和 (i^0, j^*) 也是对策的鞍点,且

$$a_{i^*j^*} = a_{i^0j^0} = a_{i^*j^0} = a_{i^0j^*}.$$

这个定理说明了具有鞍点的矩阵对策的两个性质:一是鞍点的可交换性,或称为矩形性质;二是鞍点处的值都相等.

1.2　混合策略

1.2.1　混合策略及其性质

若 $m \times n$ 矩阵对策 $A = [a_{ij}]$ 没有鞍点,即当

$$\max_{1\leqslant i\leqslant m}\min_{1\leqslant j\leqslant n} a_{ij} < \min_{1\leqslant j\leqslant n}\max_{1\leqslant i\leqslant m} a_{ij}$$

时,对策的值和解不存在.下面引入混合策略的概念.

为了避免让对方猜出自己采用哪一个策略,局中人1可以用一种随机的方法来决定自己要选择的策略,也就是采用一个混合策略.

定义2　局中人1的**混合策略**是一个 m 维向量 $X = (x_1, x_2, \cdots, x_m)$,它满足

$$x_i \geqslant 0 \quad (i = 1,2,\cdots,m) \text{ 和 } \sum_{i=1}^{m} x_i = 1.$$

以 S_m 记全体 X 的集.同样,局中人2的混合策略是一个 n 维向量 $Y = (y_1, y_2, \cdots, y_n)$,满足

$$y_j \geqslant 0 \qquad (j = 1,2,\cdots,n) \text{ 和 } \sum_{j=1}^{n} y_j = 1.$$

以 S_n 记全体 Y 的集.

局中人1以概率 x_i 选择策略 i,局中人2以概率 y_j 选择策略 j.这时支付为 a_{ij} 的概率是 x_iy_j.每一个支付 a_{ij} 乘以相应的概率 x_iy_j,对所有的 i 和所有的 j 求和,就得到局中人1的**期望支付**:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_iy_j = XAY^{\mathrm{T}}. \tag{1-7}$$

局中人 1 希望这个期望支付越大越好，局中人 2 则希望它越小越好．如果局中人 1 选择策略 $X \in S_m$，他的期望支付至少为

$$\min_{Y \in S_n} XAY^{\mathrm{T}}.$$

局中人 1 可以选择 $X \in S_m$ 使上式为最大，即可以使自己得到的支付不小于

$$v_1 = \max_{X \in S_m} \min_{Y \in S_n} XAY^{\mathrm{T}}. \tag{1-8}$$

如果局中人 2 选用策略 $Y \in S_n$，他应当付出的期望支付最多为

$$\max_{X \in S_m} XAY^{\mathrm{T}}.$$

局中人 2 可以选择 $Y \in S_n$ 使上式为最小，即他可以使局中人 1 得到的支付不大于

$$v_2 = \min_{Y \in S_n} \max_{X \in S_m} XAY^{\mathrm{T}}. \tag{1-9}$$

v_1 和 v_2 之间有下述关系．

定理 3　对于任意 $m \times n$ 矩阵对策 $A = [a_{ij}]$，有

$$v_1 = \max_{X \in S_m} \min_{Y \in S_n} XAY^{\mathrm{T}} \leqslant \min_{Y \in S_n} \max_{X \in S_m} XAY^{\mathrm{T}} = v_2. \tag{1-10}$$

1.2.2　最小最大值定理

冯·诺伊曼首先证明：对于一切 $m \times n$ 矩阵对策 $A = [a_{ij}]$，(1-8) 式和 (1-9) 式中的两个值 v_1 和 v_2 存在且相等．这一结果就是著名的对策论基本定理，或称**最小最大值定理**．

定理 4（冯·诺伊曼定理）　对于任意 $m \times n$ 矩阵对策 $A = [a_{ij}]$，有

$$\max_{X \in S_m} \min_{Y \in S_n} XAY^{\mathrm{T}} = \min_{Y \in S_n} \max_{X \in S_m} XAY^{\mathrm{T}}. \tag{1-11}$$

1.2.3　混合策略下的鞍点

(1-6) 式是纯策略下鞍点的定义．对于混合策略，也有类似的鞍点的概念．

定义 3　设 $X^* \in S_m$，$Y^* \in S_n$．如果对于一切 $X \in S_m$ 和一切 $Y \in S_n$，有

$$XAY^{*\mathrm{T}} \leqslant X^* AY^{*\mathrm{T}} \leqslant X^* AY^{\mathrm{T}}, \tag{1-12}$$

则称 (X^*, Y^*) 是 $m \times n$ 矩阵对策 $A = [a_{ij}]$ 的一个**鞍点**（在混合策略意义下）．

下面的定理表明了鞍点存在和最小最大值定理的等价性．

定理 5　$m \times n$ 矩阵对策 $A = [a_{ij}]$ 有鞍点的充要条件是 (1-8) 式和 (1-9) 式存在且相等．

(1-11) 式等号左右两边可以改写成下列等价的形式：

$$\max_{X \in S_m} \min_{Y \in S_n} \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} a_{ij} x_i \Big) y_j = \max_{X \in S_m} \min_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} a_{ij} x_i,$$

$$\min_{Y \in S_n} \max_{X \in S_m} \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} y_j \Big) x_i = \min_{Y \in S_n} \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} a_{ij} y_j.$$

于是最小最大值定理 (1-11) 式就可以改写成下面较简单实用的形式：

$$\max_{X\in S_m}\min_{1\leqslant j\leqslant n}\sum_{i=1}^{m}a_{ij}x_i = \min_{Y\in S_n}\max_{1\leqslant i\leqslant m}\sum_{j=1}^{n}a_{ij}y_j. \tag{1-13}$$

1.3 最优策略及其性质

1.3.1 最优策略

设 $m\times n$ 矩阵对策的支付矩阵是 $A=[a_{ij}]$. 由定理4和定理5可知，鞍点必定存在. 设 (X^*,Y^*) 是一个鞍点，即对于一切 $X\in S_m$ 和一切 $Y\in S_n$，有

$$XAY^{*\mathrm{T}}\leqslant X^*AY^{*\mathrm{T}}\leqslant X^*AY^{\mathrm{T}}.$$

称 X^* 和 Y^* 分别是局中人1和局中人2的最优策略，并称 $v=X^*AY^{*\mathrm{T}}$ 为对策的值. 也称 (X^*,Y^*) 为对策的一个解.

由鞍点的定义(1-12)式可知，只要局中人1坚持采用他的最优策略 X^*，则不论局中人2选择什么策略，局中人1的期望支付不会少于对策的值.

$$v=X^*AY^{*\mathrm{T}}.$$

同样，只要局中人2坚持采用他的最优策略 Y^*，则不论局中人1选择什么策略，都不能使期望支付超过对策的值 v.

1.3.2 最优策略的性质

为了方便，下面采用一些常见的矩阵记号.

以 $A_{i\cdot}$ 表示矩阵 A 第 i 个行向量，以 $A_{\cdot j}$ 表示它的第 j 个列向量，则

$$XA_{\cdot j}=\sum_{i=1}^{m}a_{ij}x_i,$$

$$A_{i\cdot}Y^{\mathrm{T}}=\sum_{j=1}^{n}a_{ij}y_j.$$

第一个式子表示局中人1采用混合策略 X，而局中人2采用纯策略 j 时的期望支付；第二个式子表示局中人2采用混合策略 Y，而局中人1采用纯策略 i 时的期望支付.

下面是最优策略的两个最重要的性质.

设 $m\times n$ 矩阵对策 $A=[a_{ij}]$ 的值是 v.

定理6 设 Y^* 是局中人2的一个最优策略. 如果对于某个 i，有

$$A_{i\cdot}Y^{*\mathrm{T}}<v,$$

则在局中人1的任一个最优策略 X^* 中必有 $x_i^*=0$.

定理7 设 X^* 是局中人1的一个最优策略. 如果对于某个 j，有

$$X^*A_{\cdot j}>v,$$

则在局中人2的任一个最优策略 Y^* 中必有 $y_j^*=0$.

由定理6可知，当已知对策的值是 v，并且 Y^* 是局中人2的一个最优策略时，若局中人1采用纯策略 i 时他的期望支付达不到 v，则 i 这个纯策略是不可取的，在局中人1的任何一个最优策略 X^* 中一定不会包含这个纯策略.

定理 7 的意义与此类似.

定理 8 $X^* \in S_m$ 是局中人 1 的最优策略的充要条件为

$$X^* A_{\cdot j} \geqslant v \quad (j = 1,2,\cdots,n).$$

定理 9 $Y^* \in S_n$ 是局中人 2 的最优策略的充要条件为

$$A_{i\cdot} Y^{*\mathrm{T}} \leqslant v \quad (i = 1,2,\cdots,m).$$

如果已知对策的值 v,就可以利用这两个定理检验局中人 1 或局中人 2 的某个策略是否是他的最优策略.

1.4 策略的优超关系

设矩阵对策的支付矩阵为

$$\begin{bmatrix} 0 & 1 & -1 \\ 1 & -2 & 0 \\ 2 & -1 & 1 \end{bmatrix}.$$

对支付矩阵的元素稍加考察,就能看出,局中人 1 决不会采用他的第 2 个策略. 这是因为,不论局中人 2 选择什么策略,局中人 1 的第 3 个策略得到的支付总比第 2 个策略得到的支付为大. 因此,局中人 1 的第 2 个策略只能以零概率出现在他的最优混合策略中.

于是,要解上面这个矩阵对策,可以将矩阵的第 2 行划去,只要解矩阵对策

$$\begin{bmatrix} 0 & 1 & -1 \\ 2 & -1 & 1 \end{bmatrix}$$

就行了. 而在这个 2×3 矩阵对策中,局中人 2 显然不愿采用策略 1:不论局中人 1 采用哪一个策略,第 1 列的支付总比第 3 列的支付大. 因此,可以将这个矩阵对策的第 1 列划去,只要解矩阵对策

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

就行了. 容易验证,这个 2×2 矩阵对策的解是 $X^* = (1/2,1/2)$, $Y^* = (1/2,1/2)$.

回到原来的 3×3 矩阵对策,显然它的解应是 $X^* = (1/2,0,1/2)$, $Y^* = (0,1/2,1/2)$, $v = 0$.

下面介绍策略间优超关系的概念.

定义 4 设 $A = [a_{ij}]$ 是 $m\times n$ 矩阵对策. 如果 $a_{kj} \geqslant a_{lj}$, $j = 1,2,\cdots,n$,则称局中人 1 的策略 k 优超于策略 l. 如果 $a_{ik} \leqslant a_{il}$, $i = 1,2,\cdots,m$,则称局中人 2 的策略 k 优超于策略 l.

如果在以上两组式子里成立严格的不等号,则分别称局中人 1、局中人 2 的策略 k **严格优超**于策略 l.

从上面的例子可以看出,利用优超关系可以简化求解矩阵对策的过程.

对于混合策略,也有类似的优超关系概念.

有时,还可以考虑一个纯策略被另外几个纯策略的凸线性组合所优超的情形. 在这种情形下,如果是严格优超,则将被优超的那个纯策略所对应的行或列划去

后,由剩下的较小的矩阵对策的最优策略,立即得到原来对策的最优策略,这只要将划去的那一行或列所对应的纯策略赋以零概率即可.

如果是优超而不是严格优超,则仍可以从较小的矩阵对策的解得到原来对策的解,但这时有可能会"失去"某些解.这就是说,从较小矩阵对策的最优策略,通过加上零概率的办法得到原来对策的最优策略时,所得到的可能不是全部解.但在通常的情形下,往往只要求得到一个解,而不一定要得到全部解,这时就可以应用这种优超关系求解.

1.5　2×2 矩阵对策的解

考虑 2×2 矩阵对策

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{1-14}$$

如果对策有纯策略的鞍点,则立即可得到纯策略解.

在没有鞍点的情况下,通过两行或两列的互换,也就是通过局中人 1 或局中人 2 的两个策略的编号的互换,或者通过矩阵的转置和各元素的变号,也就是通过局中人 1、局中人 2 名称的互换,不难发现,只须考虑以下情形:

$$a < b,\quad a < c,\quad d < b,\quad d < c.$$

这时对策必有混合策略解.

应用定理 6、定理 7,可以求得(1-14)式中对策的解为

$$X^* = (x^*, 1 - x^*),$$
$$Y^* = (y^*, 1 - y^*).$$

其中

$$x^* = \frac{d - c}{a + d - b - c};\quad y^* = \frac{d - b}{a + d - b - c}. \tag{1-15}$$

对策的值为

$$v = \frac{ad - bc}{a + d - b - c}. \tag{1-16}$$

这些公式对于

$$a > b,\quad a > c,\quad d > b,\quad d > c$$

的情形同样适用.

1.6　2×n 和 m×2 矩阵对策的图解法

对于 2×n 的情形,以 $n = 3$ 为例,设 2×3 矩阵对策的支付矩阵 A 为

$$\begin{array}{ccc} & & \boxed{1}\ \ \boxed{2}\ \ \boxed{3} \\ x & ① & \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \\ 1 - x & ② & \end{array}$$

以 ①,② 分别表示局中人 1 的第 1,2 个纯策略;$\boxed{1}$,$\boxed{2}$,$\boxed{3}$分别表示局中人 2 的

第 1,2,3 个纯策略.

设局中人 1 采用混合策略

$$X = (x_1, x_2) = (x, 1 - x),$$

其中 $0 \leqslant x \leqslant 1$. $x = 1$ 代表第 1 个纯策略 ①, $x = 0$ 代表第 2 个纯策略 ②.

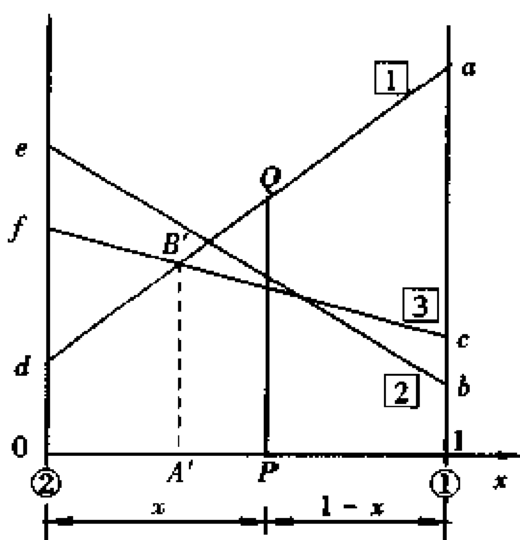


图 1-1

当 $x = 1$, 即局中人 1 采用纯策略 ① 时, 若局中人 2 采用纯策略 [1], 支付为 a, 如图 1-1 所示. 当 $x = 0$, 即局中人 1 采用策略 ② 时, 对应于 [1] 的支付为 d. 连接图中 a, d 点形成的直线 ad 表示期望支付.

设 x 轴上点 P 的坐标是 x. 容易证明, 纵坐标 PQ 是局中人 1 采用混合策略 X 而局中人 2 采用纯策略 [1] 时的期望支付, 即 $XA_{\cdot 1}$.

同样, be 和 cf 上的点的纵坐标分别表示局中人 1 采用 X, 而局中人 2 采用 [2] 和 [3] 时的期望支付.

对于每一个混合策略 X, 局中人 1 至少可以得到 ad, be, cf 三条直线在 x 处纵坐标的最小值, 即

$$\min_{1 \leqslant j \leqslant 3} XA_{\cdot j} = \min_{1 \leqslant j \leqslant 3} \sum_{i=1}^{2} a_{ij} x_i .$$

图 1-1 所示的粗折线 $dB'b$ 表示这个最小值函数.

局中人 1 希望选择 X 使上面这个最小值尽可能地大. 从图 1-1 中可以看出来, 他应当选择点 A' 所代表的 X, 这时上述的最小值为最大, 即

$$A'B' = \max_{X \in S_2} \min_{1 \leqslant j \leqslant 3} XA_{\cdot j} .$$

由 (1-13) 式可知, 这就是对策的值.

在图 1-1 中, 点 B' 是 ad 和 cf 两条直线的交点. 只要解两个二元一次联立方程, 就可求出 A' 的坐标 $x = x^*$ 和 $A'B'$ 的值. 由图 1-1 也可看出, 这里局中人 2 的最优策略不涉及他的纯策略 [2].

这种图解法可以应用于一切 $2 \times n$ 的矩阵对策.

$m \times 2$ 矩阵对策的图解法与此类似.

1.7 矩阵对策与线性规划的关系

设 $A = [a_{ij}]$ 是 $m \times n$ 矩阵对策的支付矩阵. 可假设对于一切 i 和一切 j, 有 $a_{ij} > 0$, 则对策的值 $v > 0$; 否则只须对 A 的每个元素加上一个适当的正的常数, 就可满足上述条件.

当局中人 1 采用混合策略 $X \in S_m$ 时, 他的期望支付至少为

因此，
$$\min_{1\leqslant j\leqslant n} XA_{\cdot j} = u.$$
$$XA_{\cdot j} \geqslant u \quad (j = 1,2,\cdots,n),$$
即

$$\begin{cases} \sum_{i=1}^{m} a_{ij}x_i \geqslant u & (j = 1,2,\cdots,n), \\ \sum_{i=1}^{m} x_i = 1, \\ x_i \geqslant 0 & (i = 1,2,\cdots,m). \end{cases}$$

令 $x_i/u = x'_i (i = 1,2,\cdots,m)$，则上面各式化为

$$\begin{cases} \sum_{i=1}^{m} a_{ij}x'_i \geqslant 1 & (j = 1,2,\cdots,n), \\ \sum_{i=1}^{m} x'_i = \frac{1}{u}, \\ x'_i \geqslant 0 & (i = 1,2,\cdots,m). \end{cases}$$

由于局中人1希望使 u 为极大(根据(1-13)式，这个极大值就是对策的值 v)，也就是使 $1/u$ 为极小，因此，上述问题可化为下列线性规划问题：

$$\begin{cases} \min & x'_1 + x'_2 + \cdots + x'_m, \\ \text{s.t.} & \sum_{i=1}^{m} a_{ij}x'_i \geqslant 1 \quad (j = 1,2,\cdots,n), \\ & x'_i \geqslant 0 \quad (i = 1,2,\cdots,m). \end{cases} \tag{1-17}$$

类似地，当局中人2采用混合策略 $Y \in S_n$ 时，局中人1得到的期望支付不超过
$$\max_{1\leqslant i\leqslant m} A_{i\cdot} Y^{\mathrm{T}} = w.$$
因此，
$$A_{i\cdot} Y^{\mathrm{T}} \leqslant w \quad (i = 1,2,\cdots,m),$$
即

$$\begin{cases} \sum_{j=1}^{n} a_{ij}y_j \leqslant w & (i = 1,2,\cdots,m), \\ \sum_{j=1}^{n} y_j = 1, \\ y_j \geqslant 0 & (j = 1,2,\cdots,n). \end{cases}$$

令 $y_j/w = y'_j (j = 1,2,\cdots,n)$. 由于局中人2希望使 w 为极小(这个极小值也是 v)，也就是使 $1/w$ 为极大，因此，上述问题可化为下列线性规划问题：

$$\begin{cases} \max & y'_1 + y'_2 + \cdots + y'_n, \\ \text{s.t.} & \sum_{i=1}^{m} a_{ij}y'_j \leqslant 1 \quad (i = 1,2,\cdots,m), \\ & y'_j \geqslant 0 \quad (j = 1,2,\cdots,n). \end{cases} \tag{1-18}$$

(1-17)式和(1-18)式是对偶线性规划问题. $m \times n$ 矩阵对策 $A = [a_{ij}]$ 的求解

问题等价于求解上述对偶线性规划问题.

2　无限对策

2.1　零和二人无限对策

定义 1　局中人 1 从区间[0,1]中选择一个数 x,局中人 2 完全独立地从区间[0,1]中选择一个数 y. x 和 y 称为局中人 1、局中人 2 的纯策略. 选定 x,y 后,局中人 1 得到支付 $P(x,y)$,局中人 2 得到支付 $-P(x,y)$. 这种对策称为**零和二人无限对策**,或称为正方形上的无限对策.

同矩阵对策的情形一样,下面的不等式必定成立:

$$\max_{0\leqslant x\leqslant 1}\min_{0\leqslant y\leqslant 1}P(x,y)\leqslant\min_{0\leqslant y\leqslant 1}\max_{0\leqslant x\leqslant 1}P(x,y), \tag{2-1}$$

假定两端的值都存在.

如果

$$\max_{0\leqslant x\leqslant 1}\min_{0\leqslant y\leqslant 1}P(x,y)=\min_{0\leqslant y\leqslant 1}\max_{0\leqslant x\leqslant 1}P(x,y), \tag{2-2}$$

则存在点$(x^*,y^*)\in[0,1]\times[0,1]$,使得不等式

$$P(x,y^*)\leqslant P(x^*,y^*)\leqslant P(x^*,y)$$

对于一切 $x\in[0,1]$ 和一切 $y\in[0,1]$ 成立. 这时,称(x^*,y^*)为支付函数 $P(x,y)$ 或对策的一个(纯策略)鞍点. $P(x,y)$ 在鞍点处的值

$$v=P(x^*,y^*)$$

称为对策的值. 且有

$$\max_{0\leqslant x\leqslant 1}\min_{0\leqslant y\leqslant 1}P(x,y)=P(x^*,y^*)=\min_{0\leqslant y\leqslant 1}\max_{0\leqslant x\leqslant 1}P(x,y), \tag{2-3}$$

$$\max_{0\leqslant x\leqslant 1}P(x,y^*)=P(x^*,y^*)=\min_{0\leqslant y\leqslant 1}P(x^*,y). \tag{2-4}$$

2.2　混合策略

如果(2-2)式不成立,则(2-1)式中的严格不等号成立,即

$$\max_{0\leqslant x\leqslant 1}\min_{0\leqslant y\leqslant 1}P(x,y)<\min_{0\leqslant y\leqslant 1}\max_{0\leqslant x\leqslant 1}P(x,y).$$

这时,就需要引进混合策略的概念.

定义 2　正方形上无限对策局中人 1 的混合策略是定义在[0,1]上的一个分布函数 $F(x)$:对于每一个 $x\in[0,1]$,$F(x)$ 是用某种随机的方法选出的数小于或等于 x 的概率,也就是随机变量 ξ 的值小于或等于 x 的概率

$$F(x)=P\{\xi\leqslant x\}.$$

当 $x=0$ 时,定义

$$F(0)=P\{\xi<0\}=0.$$

由定义 2,有

$$F(b) - F(a) = P\{a < \xi \leqslant b\},$$

$$F(b) - F(0) = P\{0 \leqslant \xi \leqslant b\}.$$

局中人 2 的混合策略 $G(y)$ 也是定义在 $[0,1]$ 上的一个分布函数.

如果局中人 2 采用纯策略 y,局中人 1 采用混合策略 $F(x)$,则局中人 1 的期望支付为

$$E(F,y) = \int_0^1 P(x,y)\mathrm{d}F(x),$$

这里的积分是斯蒂尔切斯(T.J.Stieltjes)积分.

同样,如果局中人 1 采用纯策略 x,局中人 2 采用混合策略 $G(y)$,则局中人 1 的期望支付为

$$E(x,G) = \int_0^1 P(x,y)\mathrm{d}G(y).$$

如果局中人 1、局中人 2 分别采用混合策略 $F(x)$, $G(y)$,则局中人 1 的期望支付为

$$E(F,G) = \int_0^1\int_0^1 P(x,y)\mathrm{d}F(x)\mathrm{d}G(y).$$

对于任意的 F 和 G,有

$$v_1 = \max_F \min_G E(F,G) \leqslant \min_G \max_F E(F,G) = v_2. \tag{2-5}$$

这里的最小值和最大值都是在全体分布函数的集合上取的,并且假定 v_1 和 v_2 都存在.

2.3　连续对策

在一般情形下,(2-5) 式中的等号不一定成立.

定理 1　设无限对策的支付函数 $P(x,y)$ 是定义在 $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$ 上的连续函数,则(2-5) 式中的 v_1 和 v_2 存在且相等.

定义 3　支付函数是连续函数的无限对策称为连续对策.

定义 4　设 $P(x,y)$ 是 $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$ 上连续对策的支付函数.如果存在局中人 1、局中人 2 的混合策略 $F^*(x)$, $G^*(y)$,使得不等式

$$E(F,G^*) \leqslant E(F^*,G^*) \leqslant E(F^*,G)$$

对于一切分布函数 F 和 G 成立,则称 (F^*,G^*) 为 $E(F,G)$ 或连续对策的一个(混合策略下的)鞍点,或称为对策的一个解. $F^*(x)$, $G^*(y)$ 分别称为局中人 1、局中人 2 的最优(混合)策略.

定理 2　连续对策鞍点存在与定理 1 中 $v_1 = v_2$ 等价.

设 $f(x)$, $g(y)$ 分别为 $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$ 上的连续函数,则有

$$\max_F \int_0^1 f(x)\mathrm{d}F(x) = \max_{0 \leqslant x \leqslant 1} f(x), \tag{2-6}$$

$$\min_G \int_0^1 g(y)\mathrm{d}G(y) = \min_{0 \leqslant y \leqslant 1} g(y). \tag{2-7}$$

利用(2-6)式和(2-7)式可将关于连续对策的基本定理(即定理1)写成下列与之等价的形式:

$$v_1 = \max_F \min_{0\leq y\leq 1}\int_0^1 P(x,y)\mathrm{d}F(x)$$
$$= \min_G \max_{0\leq x\leq 1}\int_0^1 P(x,y)\mathrm{d}G(y) = v_2,$$

或

$$\max_{0\leq x\leq 1}\int_0^1 P(x,y)\mathrm{d}G^*(y) = E(F^*,G^*)$$
$$= \min_{0\leq y\leq 1}\int_0^1 P(x,y)\mathrm{d}F^*(x),$$

其中 $F^*(x)$, $G^*(y)$ 分别是局中人1、局中人2的最优策略.

2.4 具凸支付函数的连续对策

如果单位正方形上连续对策的支付函数 $P(x,y)$ 对于其中一个变量来说是个凸函数,则这种对策称为具凸支付函数的对策,它的求解比较简单,解法包含在下述三个定理中.

定理3 设单位正方形上连续对策的支付函数 $P(x,y)$ 对于每一个 x 是 y 的严格凸函数,则局中人2有一个最优纯策略,并且这个纯策略是局中人2的唯一最优策略.

定理4 在定理3的假设条件下,对策的值为

$$v = \min_{0\leq y\leq 1}\max_{0\leq x\leq 1} P(x,y).$$

由定理4可知,局中人2的最优纯策略 y^* 满足

$$v = \min_{0\leq y\leq 1}\max_{0\leq x\leq 1} P(x,y) = \max_{0\leq x\leq 1} P(x,y^*).$$

定理5 在定理3的假设条件下,局中人1的最优策略为 $F^*(x)$.

1° 若 $y^* = 0$,则

$$F^*(x) = I_{x^*}(x),$$

其中 $I_{x^*}(x)$ 是具有一个阶梯的阶梯分布函数,即

$$I_{x^*}(x) = \begin{cases} 0 & (0\leq x < x^*), \\ 1 & (x^* \leq x \leq 1), \end{cases}$$

$x^* \in [0,1]$,满足条件

$$\begin{cases} P(x^*,0) = v, \\ \dfrac{\partial}{\partial y}P(x^*,0) \geq 0. \end{cases}$$

2° 若 $y^* = 1$,则

$$F^*(x) = I_{x^*}(x),$$

其中 $x^* \in [0,1]$,满足条件

$$\begin{cases} P(x^*,1) = v, \\ \dfrac{\partial}{\partial y}P(x^*,1) \leqslant 0. \end{cases}$$

3° 若 $0 < y^* < 1$,则

$$F^*(x) = \alpha I_{x_1^*}(x) + (1-\alpha) I_{x_2^*}(x) \quad (0 \leqslant \alpha \leqslant 1),$$

其中 $x_1^* \in [0,1]$,$x_2^* \in [0,1]$ 和 α 满足条件:

$$\begin{cases} P(x_1^*,y^*) = v, \\ \dfrac{\partial}{\partial y}P(x_1^*,y^*) \geqslant 0, \\ P(x_2^*,y^*) = v, \\ \dfrac{\partial}{\partial y}P(x_2^*,y^*) \leqslant 0, \\ \alpha \dfrac{\partial}{\partial y}P(x_1^*,y^*) + (1-\alpha)\dfrac{\partial}{\partial y}P(x_2^*,y^*) = 0. \end{cases}$$

如果支付函数 $P(x,y)$ 是 y 的凸函数,而不是严格凸函数,则以上三个定理仍成立,但这时局中人 2 的最优策略通常不再是唯一的了.

3　非合作对策

3.1　基本概念

定义 1　称

$$\Gamma = [I, \{S_i\}, \{P_i\}]$$

为 n 人非合作对策.其中,$I = \{1,2,\cdots,n\}$ 是局中人的集;每个局中人 $i = 1,2,\cdots,n$ 有一个纯策略的有限集

$$S_i = \{s^{(i)}\} = \{s_1^{(i)}, s_2^{(i)}, \cdots, s_{m_i}^{(i)}\},$$

其中 $s_1^{(i)}, s_2^{(i)}, \cdots, s_{m_i}^{(i)}$ 是局中人 i 的 m_i 个纯策略;每个局中人 i 有一个支付函数 P_i.当每个局中人 i 选定一个策略 $s^{(i)}$ 后,就形成了对策的一个纯策略局势

$$s = (s^{(1)}, s^{(2)}, \cdots, s^{(n)}) \qquad (s^{(i)} \in S_i);$$

支付函数就是局势的函数:

$$P_i = P_i(s) \qquad (i = 1,2,\cdots,n).$$

定义 2　定义

$$s \,\|\, t^{(i)} = (s^{(1)}, s^{(2)}, \cdots, s^{(i-1)}, t^{(i)}, s^{(i+1)}, \cdots, s^{(n)}).$$

这就是在局势 $s = (s^{(1)}, s^{(2)}, \cdots, s^{(n)})$ 中,第 i 个局中人将他的策略 $s^{(i)}$ 换成 $t^{(i)}$,其他局中人的策略不变而得到的新的局势.显然,$s \,\|\, s^{(i)} = s$.

定义 3　设 s^* 是 n 人非合作对策

$$\Gamma = [I, \{S_i\}, \{P_i\}]$$

的一个局势.如果对于每一个 $i \in I$ 和每一个 $s^{(i)} \in S_i(s^{(i)} = s_k^{(i)}, k = 1,2,\cdots, m_i)$,有

$$P_i(s^* \parallel s^{(i)}) \leqslant P_i(s^*),$$

则称 s^* 是 Γ 的一个平衡点或平衡局势.

以上三个定义都是对纯策略而言的.显然,在一个 n 人非合作对策中,平衡点不一定存在.

同矩阵对策一样,也需要考虑局中人的混合策略.

对于每一个局中人 $i \in I$,以 $x^{(i)}$ 表示 i 的一个混合策略,即

$$x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_{m_i}^{(i)}),$$

其中

$$x_k^{(i)} \geqslant 0 \quad (k = 1,2,\cdots, m_i),$$

$$\sum_{k=1}^{m_i} x_k^{(i)} = 1.$$

局中人 i 以概率 $x_k^{(i)}$ 选择策略 $s_k^{(i)}, k = 1,2,\cdots, m_i$.

定义 4　称

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})$$

为对策的一个(混合策略)局势,其中 $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_{m_i}^{(i)})$ 是局中人 i 的一个混合策略,$i = 1,2,\cdots, n$.

定义 5　定义

$$x \parallel z^{(i)} = (x^{(1)}, x^{(2)}, \cdots, x^{(i-1)}, z^{(i)}, x^{(i+1)}, \cdots, x^{(n)}).$$

这就是在混合策略局势 x 中,局中人 i 将他的混合策略 $x^{(i)}$ 换成另一个混合策略 $z^{(i)}$,其他局中人的策略不变而得到的新的局势.显然,$x \parallel x^{(i)} = x$.

定义 6　称

$$\Gamma = [I, \{X_i\}, \{P_i\}]$$

为 n 人非合作对策(在混合策略意义下),其中

$$I = \{1,2,\cdots, n\},$$

$$\{X_i\} = \{X_1, X_2, \cdots, X_n\},$$

$$X_i = \{x^{(i)}\} = \{(x_1^{(i)}, x_2^{(i)}, \cdots, x_{m_i}^{(i)})\} \quad (i = 1,2,\cdots, n),$$

$$x_k^{(i)} \geqslant 0 \quad (k = 1,2,\cdots, m_i),$$

$$\sum_{k=1}^{m_i} x_k^{(i)} = 1,$$

$$\{P_i\} = \{P_1, P_2, \cdots, P_n\},$$

$$P_i = P_i(s) \quad (i = 1,2,\cdots, n).$$

为了方便,在定义1和定义6中用了同一个字母 Γ 表示纯策略和混合策略意义下的对策.由于混合策略是更一般的情况,这样做不会引起混淆.

定义7 设 x^* 是 n 人非合作对策

$$\Gamma = [I, \{X_i\}, \{P_i\}]$$

的一个混合策略局势. 以 $E_i(x)$ 表示局中人 i 在局势 x 下应得的期望支付, $i = 1, 2, \cdots, n$. 如果对于每一个 $i \in I$ 和每一个 $x^{(i)} \in X_i$, 有

$$E_i(x^* \parallel x^{(i)}) \leqslant E_i(x^*),$$

则称 x^* 是 Γ(在混合策略下) 的一个平衡点或平衡局势.

3.2 平衡点的存在性

纳什(J.F.Nash) 证明了混合策略下的平衡点必定存在.

定理1 (纳什定理) 每一个 n 人非合作对策 $\Gamma = [I, \{X_i\}, \{P_i\}]$ 必有平衡点.

3.3 2×2双矩阵对策的平衡点

定义8 设局中人1有 m 个策略 $i = 1, 2, \cdots, m$, 局中人2有 n 个策略 $j = 1, 2, \cdots, n$. 设局中人1选择策略 i, 局中人2选择策略 j 时, 他们得到的支付分别是 a_{ij} 和 b_{ij}, 则双方的支付矩阵分别为

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix},$$

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ b_{21} & & b_{2n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}.$$

对策由矩阵 A, B 完全确定. 这种对策叫做 $m \times n$ **双矩阵对策**.

以下只介绍 2×2 双矩阵对策平衡点的求法.

定理2 设 2×2 双矩阵对策局中人1、局中人2的支付矩阵分别为

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

$$B = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}.$$

以 $X = (x, 1-x)$, $Y = (y, 1-y)$ 分别表示局中人1、局中人2的混合策略, 其中 $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$.

令

$$Q = a - b - c + d, \quad q = d - b, \tag{3-1}$$

$$R = a' - b' - c' + d', \quad r = d' - c', \tag{3-2}$$

则对策的平衡点根据不同的 Q,q,R,r 值,由下面的1°和2°两组不等式确定:

1° 当 $Q=0$ 且 $q=0$ 时,有

$$0\leqslant x\leqslant 1,\quad 0\leqslant y\leqslant 1. \tag{3-3}$$

当 $Q=0,q>0$ 时,有

$$x=0,\quad 0\leqslant y\leqslant 1. \tag{3-4}$$

当 $Q=0,q<0$ 时,有

$$x=1,\quad 0\leqslant y\leqslant 1. \tag{3-5}$$

当 $Q>0$ 时,有

$$\begin{cases} x=0, & y\leqslant \dfrac{q}{Q}, \\ 0<x<1, & y=\dfrac{q}{Q}, \\ x=1, & y\geqslant \dfrac{q}{Q}. \end{cases} \tag{3-6}$$

当 $Q<0$ 时,(3-6)式右边第1和第3个关于 y 的不等式换成反方向的不等式.

2° 当 $R=0$ 且 $r=0$ 时,有

$$0\leqslant x\leqslant 1,\quad 0\leqslant y\leqslant 1. \tag{3-7}$$

当 $R=0,r>0$ 时,有

$$0\leqslant x\leqslant 1,\quad y=0. \tag{3-8}$$

当 $R=0,r<0$ 时,有

$$0\leqslant x\leqslant 1,\quad y=1. \tag{3-9}$$

当 $R>0$ 时,有

$$\begin{cases} x\leqslant \dfrac{r}{R}, & y=0, \\ x=\dfrac{r}{R}, & 0<y<1, \\ x\geqslant \dfrac{r}{R}, & y=1. \end{cases} \tag{3-10}$$

当 $R<0$ 时,(3-10)式左边第1和第3个关于 x 的不等式换成反方向的不等式.

将四组不等式(3-3)式至(3-6)式中满足对策条件的一组与(3-7)式至(3-10)式中满足条件的一组联立起来,即可求得与平衡点相对应的 x 和 y 值.

例1 **周末娱乐问题**.设2×2双矩阵对策的支付矩阵分别为

$$\boldsymbol{A}=\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix},\quad \boldsymbol{B}=\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

解 按照(3-1)式和(3-2)式,算得

$$Q=5>0,\qquad q=2,$$
$$R=5>0,\qquad r=3.$$

将这些数值分别代入(3-6)式和(3-10)式,得到

$$\begin{cases} x = 0, & y \leqslant \frac{2}{5}, \\ 0 < x < 1, & y = \frac{2}{5}, \\ x = 1, & y \geqslant \frac{2}{5}; \end{cases} \tag{3-11}$$

$$\begin{cases} x \leqslant \frac{3}{5}, & y = 0, \\ x = \frac{3}{5}, & 0 < y < 1, \\ x \geqslant \frac{3}{5}, & y = 1. \end{cases} \tag{3-12}$$

解这些不等式,求得对策有三个平衡点:

$$(x,y) = (0,0), \left(\frac{3}{5},\frac{2}{5}\right), (1,1).$$

用局中人1和局中人2的混合策略$(X,Y) = ((x,1-x),(y,1-y))$表示,即为

$$((0,1),(0,1)), \left(\left(\frac{3}{5},\frac{2}{5}\right),\left(\frac{2}{5},\frac{3}{5}\right)\right), ((1,0),(1,0)).$$

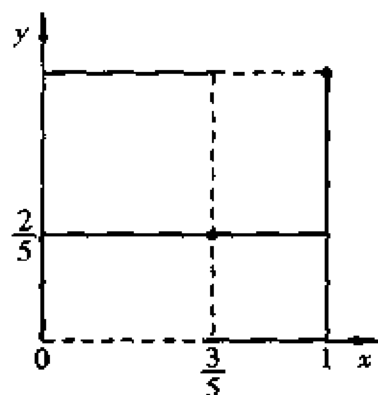


图 3-1

不等式组(3-11)式的曲线在图3-1中以粗实线条画出,(3-12)式的曲线则以虚线画出.容易看出,粗线和虚线的交点就是对策的三个平衡点.

如以(E_1,E_2)表示局中人1和局中人2的期望支付,则在上述三个平衡点处的期望支付依次为

$$(E_1,E_2)_1 = (1,2),$$

$$(E_1,E_2)_2 = \left(\frac{1}{5},\frac{1}{5}\right),$$

$$(E_1,E_2)_3 = (2,1).$$

第2个平衡点处的支付显然比第1和第3个平衡点处局中人1、局中人2所得的支付都小.但由于这是一个非合作对策,不许可在事先对如何选择策略进行协商,不许可两个局中人把他们的策略结合起来,所以无法保证达到第1或第3个平衡局势.

4　合作对策

4.1　基本概念和特征函数

在一个n人非合作对策中,两个或两个以上的局中人不许事先商定如何选择策略,不许把他们的策略结合起来.局中人之间不允许对得到的支付进行重新分

配,一个局中人不能分享另一个局中人得到的支付.

但在 n 人合作对策中,对上述两方面的问题都不加限制.局中人之间可以进行充分的合作:可以结成联盟,事先商定,把他们的策略协调结合起来;可以在终局后重新分配若干个局中人所得支付的总和.

定义 1 若 $I=\{1,2,\cdots,n\}$ 是局中人的集,$v(S)$ 是定义在 I 的一切子集即联盟 S 的集上的实值函数,并满足条件

$$v(\varnothing)=0,$$
$$v(I)\geqslant\sum_{i\in I}v(\{i\}),$$

则称 $\Gamma\equiv[I,v]$ 为 n 人**合作对策**,$v(S)$ 为对策的**特征函数**.

本章假定,每个联盟 S 得到的收入 $v(S)$ 可以按照任意方式分配给联盟的成员.这一条件称为**局外支付条件**.

定义 2 设 $\Gamma\equiv[I,v]$ 是 n 人合作对策.如果对于一切 $S,T\subset I,S\cap T=\varnothing$,有

$$v(S\cup T)\geqslant v(S)+v(T),\tag{4-1}$$

则称 v 或 Γ 具有**超可加性**.如果(4-1)式中等号恒成立,则称 v 或 Γ 具有**可加性**.具有可加性的对策称为**非实质性对策**,否则称为**实质性对策**.

定理 1 n 人合作对策 $\Gamma\equiv[I,v]$ 具有可加性的充要条件为

$$v(I)=\sum_{i\in I}v(\{i\}).$$

对于 n 人合作对策,主要要讨论的当然是实质性的对策,即特征函数 v 满足条件

$$v(I)>\sum_{i\in I}v(\{i\})$$

的对策.

4.2 策略等价关系和(0,1)规范化

n 人合作对策可以按其特征函数进行分类,以便简化对其性质的研究.

定义 3 设 $\Gamma\equiv[I,v]$ 和 $\Gamma'\equiv[I,v']$ 是定义在同一个 $I=\{1,2,\cdots,n\}$ 上的两个 n 人合作对策.如果存在 n 个常数 $a_1,a_2,\cdots,a_n$ 和一个正的常数 c,使得对于 I 的每一个子集 S,有

$$v'(S)=cv(S)+\sum_{i\in S}a_i,$$

则称 Γ 和 Γ' 是**策略等价**的,或者说,特征函数 v 和 v' 是策略等价的.

这样定义的策略等价关系显然满足等价关系的三个条件:自反性、对称性和可递性.

定义 4 若 n 人合作对策 $\Gamma\equiv[I,v]$ 的特征函数 v 满足条件

$$v(\{i\})=0\quad(i=1,2,\cdots,n)$$

和

$$v(I) = 1,$$

则称 Γ 或 v 是(0,1) **规范化的**.

定理2 每一个实质性的 n 人合作对策 $\Gamma = [I, v]$ 策略等价于唯一的一个(0,1) 规范化对策 $\Gamma' = [I, v']$. 即对于每一个 $S \subset I$,有

$$v'(S) = cv(S) + \sum_{i \in S} a_i,$$

其中

$$c = \frac{1}{v(I) - \sum_{i \in I} v(\{i\})} > 0,$$

$$a_i = \frac{-v(\{i\})}{v(I) - \sum_{i \in I} v(\{i\})} \quad (i = 1, 2, \cdots, n).$$

$v'(S)$ 是(0,1) 规范化特征函数.

4.3 二人合作对策

定义5 在一个 n 人对策中,如果对于每一个局势 s 有

$$\sum_{i=1}^{n} P_i(s) = k,$$

其中 k 为一常数,则称对策为**常和对策**,否则为**非常和对策**.

常和 n 人合作对策的特征函数有下面的性质.

定理3 设 $\Gamma = [I, v]$ 是常和 n 人合作对策,则对于每一个 $S \subset I$,有

$$v(S) + v(I \setminus S) = v(I). \tag{4-2}$$

这一性质称为**特征函数的互补性**.

最简单的合作对策是二人合作对策.如果将二人合作对策分为常和与非常和两类,则根据(4-2)式容易看出,一切常和二人合作对策都是非实质性的.

下面是一个非常和二人合作对策的例子.

例1 仍以3.3节例1为例,其中 2×2 双矩阵对策为

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

解 作为二人非合作对策,由于局中人1、局中人2只能各自独立地采用他们的混合策略,所以只有三个平衡局势.

现在,如果支付矩阵 A, B 所代表的对策是一个合作对策,则局中人1和局中人2的混合策略可以按任意方式结合起来,形成联合的混合策略.例如,两个局中人可以事先商定,或者两人都选择策略1,或者都选择策略2,以便获得更多的支付.要做到这一点,只要掷一枚钱币,预先约定,若钱币正面向上,则两人都选择策略1,若反面向上,则都选择策略2.

作为二人合作对策,只要考虑两个局中人一切可能的混合策略的组合,就可以算出相应的支付,即这个合作对策的特征函数 v 的值为

$$v(\{1\}) = v(\{2\}) = \frac{1}{5}, \quad v(\{1,2\}) = 3.$$

两个局中人只要组成联盟$\{1,2\}$,就可以获得总支付3,即$v(\{1,2\})$的值.

4.4　转归及其优超关系

定义 6　若n维向量$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$满足条件

$$x_i \geqslant v(\{i\}) \quad (i = 1,2,\cdots,n) \tag{4-3}$$

和

$$\sum_{i=1}^{n} x_i = v(I), \tag{4-4}$$

则称$\boldsymbol{x}$为对策$\Gamma = [I, v]$的一个**转归**,也可以称为分配.

(4-3)式称为**个体合理性条件**,它表示局中人i得到的分配不小于他一个人"单干"所能得到的收入.(4-4)式称为**集体合理性或帕雷托(Pareto)最优性条件**,它表示n个局中人分配之和应等于整个大联盟I的总收入.

定理 4　非实质性的n人合作对策$\Gamma = [I, v]$只有一个转归,即

$$\boldsymbol{x} = (v(\{1\}), v(\{2\}), \cdots, v(\{n\})).$$

对于实质性的对策$\Gamma = [I, v]$,由于

$$a = v(I) - \sum_{i=1}^{n} v(\{i\}) > 0,$$

故有无穷多种方式将a分为n个非负的实数$a_1, a_2, \cdots, a_n$,使得$\sum_{i=1}^{n} a_i = a$.因此,任何形如

$$\boldsymbol{x} = (v(\{1\}) + a_1, v(\{2\}) + a_2, \cdots, v(\{n\}) + a_n)$$

的向量都是Γ的转归.

对于(0,1)规范化的对策$\Gamma = [I, v]$,转归$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$应满足的条件(4-3)式、(4-4)式变为

$$x_i \geqslant 0 \qquad (i = 1,2,\cdots,n), \tag{4-5}$$

$$\sum_{i=1}^{n} x_i = 1. \tag{4-6}$$

定义 7　n人合作对策$\Gamma = [I, v]$的全体转归的集记为$X(\Gamma)$.

定义 8　设$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$和$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$是$n$人合作对策$\Gamma = [I, v]$的两个转归,联盟$S$是$I$的非空子集.如果

$$v(S) \geqslant \sum_{i \in S} y_i, \tag{4-7}$$

且

$$y_i > x_i \quad (i \in S), \tag{4-8}$$

则称$\boldsymbol{y}$关于S优超于$\boldsymbol{x}$,或者说,$\boldsymbol{x}$关于S被$\boldsymbol{y}$优超,记为$\boldsymbol{y} >_S \boldsymbol{x}$.

(4-7)式称为**有效性条件或可行性条件**.或者说,S是对于$\boldsymbol{y}$的**有效集**.

一个转归关于某个联盟优超于另一个转归的关系满足可递性.这就是说,如果

$$z >_S y, \quad y >_S x,$$

则

$$z >_S x.$$

由定义可知,关于单人联盟$\{i\}$不可能有转归的优超关系.关于全体局中人的大联盟I也不可能有转归的优超关系.

定义9　如果存在非空联盟$S \subset I$,使

$$y >_S x,$$

则称转归y优超于x,记为$y > x$.

转归的一般优超关系不一定满足可递性,举例如下.

设三人合作对策的特征函数v的值为

$$v(\{i\}) = 0 \quad (i = 1,2,3),$$

$$v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = v(\{1,2,3\}) = 6.$$

考虑转归

$$z = (0,3,3), \quad y = (4,2,0), \quad x = (2,0,4).$$

显然有$z > y, y > x$,但z不优超于x.

4.5　核　　心

为了书写简洁,采用下述记号.设$S \neq \varnothing$是一个联盟,$x = (x_1, x_2, \cdots, x_n)$是一个转归.记

$$x(S) = \sum_{i \in S} x_i,$$

并规定$x(\varnothing) = 0$.

定义10　设$\Gamma = [I, v]$是n人合作对策.定义

$$C(\Gamma) = \{x \mid x \in X(\Gamma); v(S) - x(S) \leqslant 0, S \subset I\}, \tag{4-9}$$

称之为Γ的核心.

这个定义表示,对于每一个联盟$S \subset I$,$C(\Gamma)$中的转归x提供给S的分配不少于S自身所能得到的收入$v(S)$,因而,x是能被一切S接受的转归.

定理5　设$\Gamma = [I, v]$是n人合作对策.$x \in C(\Gamma)$的充要条件是x不被优超.

单人联盟$\{i\}$和全体局中人的大联盟I都不存在转归的优超关系,所以n人合作对策只有当$n \geqslant 3$时才会有核心.

当$n \geqslant 3$时,任何非实质性的对策$\Gamma = [I, v]$具有可加性,且只有唯一的一个转归(见定理4),即

$$x = (v(\{1\}), v(\{2\}), \cdots, v(\{n\})),$$

因而这个转归也可以说就是它的核心.

$n \geqslant 3$的实质性对策,可以区分为常和n人合作对策与非常和n人合作对策两类.关于前者,有下面的定理.

定理6　设$\Gamma = [I, v]$是实质性常和n人合作对策,则

$$C(\Gamma) = \varnothing.$$

关于实质性的非常和 n 人合作对策,核心可以是非空的转归集,也可能是空集.请看 下面的例子.

首先,为了简化记号,再规定下列记法:

$$x(12) = x(\{1,2\}) = x_1 + x_2,$$
$$x(2) = x(\{2\}) = x_2,$$
$$v(23) = v(\{2,3\}), \quad v(i) = v(\{i\}),$$

等等.

例 2　设三人合作对策 Γ 的特征函数 v 的值为

$$v(i) = 0 \qquad (i = 1,2,3),$$
$$v(12) = \frac{1}{3}, \quad v(13) = \frac{1}{6},$$
$$v(23) = \frac{5}{6}, \quad v(123) = 1.$$

$x \in C(\Gamma)$ 的条件为

$$v(i) = 0 \qquad (i = 1,2,3),$$
$$v(12) = \frac{1}{3} \leqslant x_1 + x_2,$$
$$v(13) = \frac{1}{6} \leqslant x_1 + x_3,$$
$$v(23) = \frac{5}{6} \leqslant x_2 + x_3,$$
$$v(123) = 1 = x_1 + x_2 + x_3.$$

因此,

$$C(\Gamma) = \{x \mid x_1 \leqslant \frac{1}{6}, x_2 \leqslant \frac{5}{6}, x_3 \leqslant \frac{2}{3}\},$$

它是一个无穷点集.

实质性的非常和三人合作对策不一定有非空的核心.它的核心也可能是空集.事实上,对于(0,1)规范化的非常和三人合作对策来说,核心为空集的充要条件为

$$v(12) + v(13) + v(23) > 2.$$

4.6　稳　定　集

以上讨论的核心无疑是合作对策的一个重要因素,但是,如果企图以核心作为合作对策的解,却存在着不可克服的困难:有许多对策的核心是空集.本节介绍合作对策的一种古典的解.

定义 11　设 $\Gamma = [I, v]$ 是 n 人合作对策,$V \subset X(\Gamma)$ 是满足下面两个条件的转归的集:

1° 对于任意 $x, y \in V$,有 $x \not\succ y$,

2° 若 $w \in X(\Gamma)$,$w \bar{\in} V$,则存在 $z \in V$,使 $z \succ w$,

则称 V 为对策 Γ 的一个**稳定集**，或称为 Γ 的一个 **VN-M 解**（冯·诺伊曼-摩根斯顿解）.

定义中的条件 1° 表明，V 中的任意两个转归之间没有优超和被优超的关系. 这个性质称为 V 的**内部稳定性**.

条件 2° 称为 V 的**外部稳定性**. 这个性质表明，不在 V 中的每一个转归 w，至少被 V 中一个转归 z 所优超. 这就是说，至少有一个联盟 S 不喜欢 w. 这个联盟 S 为了自身的利益希望争取一个分配方案 $z \in V$，使得 $z \succ_S w$.

每一个稳定集是合作对策 Γ 在上述意义下的一个解. 这种解可以有不止一个，可以有无穷多个.

合作对策的核心和稳定集之间有下述关系.

定理 7 设 n 人合作对策 $\Gamma = [I, v]$ 有非空的核心 $C(\Gamma)$，且它的 VN-M 解 V 存在，则

$$C(\Gamma) \subset V.$$

由于稳定集往往是个无穷点集，而且已有人举出了稳定集不存在的合作对策的例子，也有核心为空集而稳定集不存在的合作对策的例子，因此，以稳定集作为合作对策的解，是不能令人满意的.

4.7 广义转归与强 ε 核心

为了后面两节的需要，这里再引进一些新的概念.

定义 12 设 $\Gamma = [I, v]$ 是 n 人合作对策. n 维向量 $x = (x_1, x_2, \cdots, x_n)$ 若满足条件 $\sum_{i \in I} x_i = v(I)$，则称为一个**广义转归**. 记全体广义转归的集为 $X^*(\Gamma)$.

Γ 的广义转归与定义 6 中转归的区别在于后者还要满足个体合理性条件(4-3)式. 显然 $X(\Gamma)$ 是 $X^*(\Gamma)$ 的子集.

定义 13 设 $\Gamma = [I, v]$ 是 n 人合作对策. 对于每一个广义转归 $x \in X^*(\Gamma)$ 和每一个联盟 $S \subset I$，定义

$$e(S, x) = v(S) - x(S),$$

称之为 S 在 x 处的**超出值**.

利用超出值的记号可将核心的定义(4-9)式改写成

$$C(\Gamma) = \{x \mid x \in X(\Gamma), e(S, x) \leqslant 0, S \subset I\}.$$

这个定义等价于

$$C(\Gamma) = \{x \mid x \in X^*(\Gamma), e(S, x) \leqslant 0, S \subset I\}.$$

定义 14 设 $\Gamma = [I, v]$ 是 n 人合作对策，ε 是一个实数. 广义转归的集

$$C_\varepsilon(\Gamma) = \{x \mid x \in X^*(\Gamma), e(S, x) \leqslant \varepsilon, S \subset I, S \neq \varnothing, I\}$$

称为 Γ 的**强 ε 核心**，或简称为 **ε 核心**.

Γ 的核心 $C(\Gamma)$ 就是它的强 0 核心 $C_0(\Gamma)$. ε 当然也可以是负数. 当 ε 足够小时，$C_\varepsilon(\Gamma) = \varnothing$；当 ε 充分大时，$C_\varepsilon(\Gamma) \neq \varnothing$. 如果 $\varepsilon_1 < \varepsilon_2$，则 $C_{\varepsilon_1}(\Gamma) \subset C_{\varepsilon_2}(\Gamma)$.

定义 15　设 $\Gamma = [I, v]$ 是 n 人合作对策. 若 ϵ_0 是使得 $C_\epsilon(\Gamma) \neq \varnothing$ 的最小 ϵ 值,则称 $C_{\epsilon_0}(\Gamma)$ 为 Γ 的最小核心,记为 $\mathrm{LC}(\Gamma)$.

$\mathrm{LC}(\Gamma) = C_{\epsilon_0}(\Gamma)$ 是 Γ 的一切非空强 ϵ 核心之交.

在下面的例 3 中,要对一个三人合作对策求出其最小核心.

假定三人合作对策已简化为(0,1)规范化的形式,则其转归集 $X(\Gamma)$ 中的点 $x = (x_1, x_2, x_3)$ 满足 $x_i \geq 0, i = 1,2,3$ 和 $x_1 + x_2 + x_3 = 1$. 可用下述方法在平面上表示这些点.

设 x 是高为 1 的等边三角形中的任意一点,如图 4-1 所示. 如果从点 x 到 1,2,3 这三个顶点的对边的距离分别为 x_1, x_2, x_3,则 x_1, x_2, x_3 满足上述转归的条件. 因此,可以 (x_1, x_2, x_3) 作为点 x 的坐标,称之为重心坐标. 顶点 1 的重心坐标是(1,0,0). 类似地,顶点 2,3 的重心坐标分别是(0,1,0)和(0,0,1). 闭三角形中全部点的集合就是全体转归的集 $X(\Gamma)$. 等边三角形的三条边 23,31,12 的方程分别是

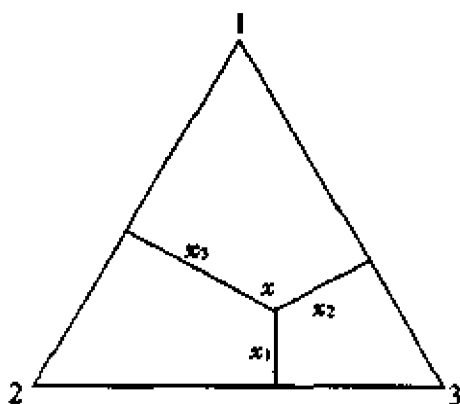


图 4-1

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 0.$$

在等边三角形外面的点,其重心坐标至少有一个为负数.

在不致引起混淆的情况下,将 $e(\{2,3\}, x)$ 简记为 $e(23)$, $e(\{1\}, x)$ 简记为 $e(1)$,等等.

例 3　前面的例 2 中三人合作对策 Γ 的特征函数 v 的值为

$$v(i) = 0 \qquad (i = 1,2,3),$$

$$v(12) = \frac{1}{3}, \quad v(13) = \frac{1}{6},$$

$$v(23) = \frac{5}{6}, \quad v(123) = 1.$$

在例 2 中已求出 Γ 的核心 $C(\Gamma) = C_0(\Gamma)$,它就是图 4-2 所示的一个四边形区域.

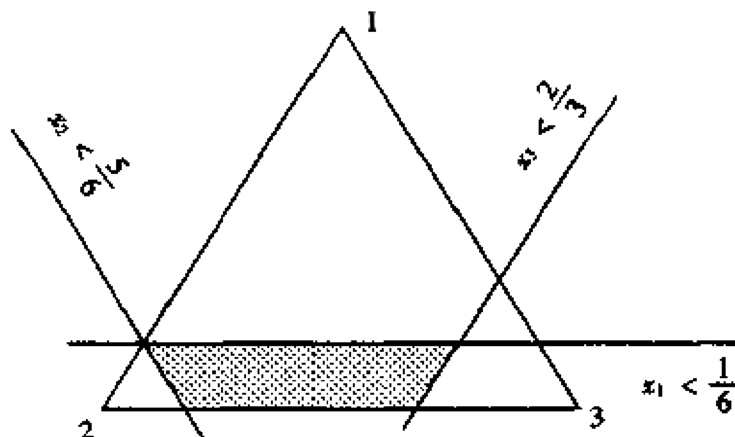


图 4-2

在图 4-3 中，除了核心 $C(\Gamma)$ 外，还画出了当 $\epsilon=\frac{1}{6}$ 和 $\epsilon=\frac{2}{6}$ 时的强 ϵ 核心 $C_\epsilon(\Gamma)$ 的图形. $C_{1/6}(\Gamma)$ 是个五边形，$C_{2/6}(\Gamma)$ 是个六边形. 最小核心是 $C_{-1/12}(\Gamma)$，它是平行于转归三角形 $X(\Gamma)$ 的边 23 的一个直线段，即

$$\mathrm{LC}(\Gamma)=C_{-1/12}(\Gamma)=\left\{x \mid x_1=\frac{1}{12}, x_2\leqslant\frac{9}{12}, x_3\leqslant\frac{7}{12}\right\}.$$

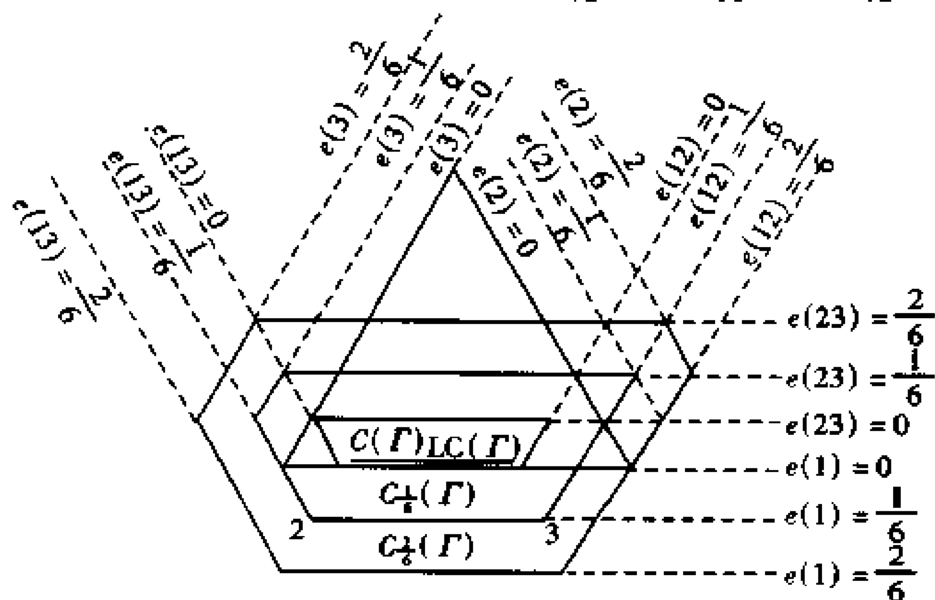


图 4-3

例 4　设三人合作对策 Γ 的特征函数 v 的值为

$$v(i)=0 \qquad (i=1,2,3),$$
$$v(12)=4, \qquad v(13)=3,$$
$$v(23)=10, \qquad v(123)=8.$$

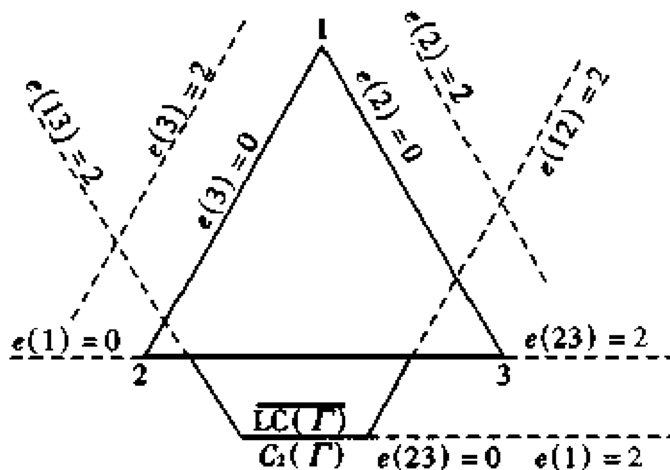


图 4-4

特征函数 v 不满足超可加性.

这个对策的核心 $C(\Gamma)$ 是空集.

在图 4-4 中画出了 Γ 的强 2 核心 $C_2(\Gamma)$ 的图形,它是由直线 $e(1)=2, e(2)=2, e(3)=2, e(12)=2, e(13)=2$ 和 $e(23)=2$ 围成的区域,是一个四边形域.

最小核心是

$$LC(\Gamma)=C_1(\Gamma)=\{x \mid x_1=-1, x_2 \leqslant 6, x_3 \leqslant 5\}.$$

图 4-4 中三角形 123 的高为 8.

4.8　核

设 $\Gamma=[I, v]$ 是 n 人合作对策.

定义 16　以 T_{ij} 表示一切包含局中人 i 而不包含局中人 j 的联盟的集,即

$$T_{ij}=\{S \mid S \subset I, i \in S, j \bar{\in} S\}.$$

例如,当 $I=\{1,2,3,4\}$ 时,

$$T_{42}=\{\{4\},\{4,1\},\{4,3\},\{4,1,3\}\}.$$

定义 17　对于每一个广义转归 $x \in X^*(\Gamma)$,定义

$$s_{ij}(x)=\max_{S \in T_{ij}} e(S, x),$$

称之为在 x 处局中人 i 超过局中人 j 的**最大剩余**.如果

$$s_{ij}(x)>s_{ji}(x),$$

则称在 x 处局中人 i 胜过局中人 j.如果在 x 处局中人 i 不胜过局中人 j,局中人 j 也不胜过局中人 i,即

$$s_{ij}(x)=s_{ji}(x), \tag{4-10}$$

则称局中人 i 和局中人 j 在 x 处平衡.

以上最大剩余、胜过与平衡的概念都是对于广义转归 $x \in X^*(\Gamma)$ 定义的.对于转归 $x \in X(\Gamma)$,最大剩余的概念完全一样.胜过与平衡的概念则有所不同.

定义 18　设 $x \in X(\Gamma)$.如果

$$s_{ij}(x)>s_{ji}(x), \quad x_j>v(j),$$

则称局中人 i 在 x 处胜过局中人 j.

如果局中人 i 不胜过局中人 j,局中人 j 也不胜过局中人 i,则称局中人 i 和局中人 j 在 x 处平衡.

由定义 18 可知,局中人 i 和局中人 j 平衡的条件为

$$(s_{ij}(x)-s_{ji}(x))(x_j-v(j)) \leqslant 0, \tag{4-11}$$

$$(s_{ji}(x)-s_{ij}(x))(x_i-v(i)) \leqslant 0. \tag{4-12}$$

定义 19　n 人合作对策 $\Gamma=[I, v]$ 的预核是广义转归 $x \in X^*(\Gamma)$ 的集 $K^*(\Gamma)$,在其中的每一个 x 处,每两个局中人 i 和 j 关于 $X^*(\Gamma)$ 平衡.

由定义 19 可知,广义转归 $x \in K^*(\Gamma)$ 的充要条件为,对于每两个局中人 i 和 j,(4-10) 式成立.

定义 20　n 人合作对策 $\Gamma=[I, v]$ 的核是转归 $x \in X(\Gamma)$ 的集 $K(\Gamma)$,在其中的每一个 x 处,每两个局中人 i 和 j 关于 $X(\Gamma)$ 平衡.

由定义 20 可知,转归 $x \in K(\Gamma)$ 的充要条件为,对于每一对局中人 i 和 j, $i \neq j$, (4-11) 式和(4-12) 式同时成立.这等价于

$$s_{ij}(x) = s_{ji}(x), \tag{4-13}$$

或

$$\begin{cases} s_{ij}(x) > s_{ji}(x), \\ x_j = v(j), \end{cases} \tag{4-14}$$

或

$$\begin{cases} s_{ji}(x) > s_{ij}(x), \\ x_i = v(i). \end{cases} \tag{4-15}$$

比较定义 18 和定义 19,预核 $K^*(\Gamma)$ 显然比核 $K(\Gamma)$ 容易计算.

对于较大的 n,合作对策核的计算一般说来是十分复杂的.以下只通过两个三人合作对策的例子加以说明.

例 5　考虑例 2 中的三人合作对策 Γ.它的特征函数 v 的值为

$$v(i) = 0 \qquad (i = 1,2,3),$$

$$v(12) = \frac{1}{3}, \qquad v(13) = \frac{1}{6},$$

$$v(23) = \frac{5}{6}, \qquad v(123) = 1.$$

为了计算预核 $K^*(\Gamma)$ 和核 $K(\Gamma)$,对于每一对局中人 i 和 j 画出方程(4-10)式的图形,如图 4-5 所示.例如,为了画出

$$s_{13}(x) = s_{31}(x) \tag{4-16}$$

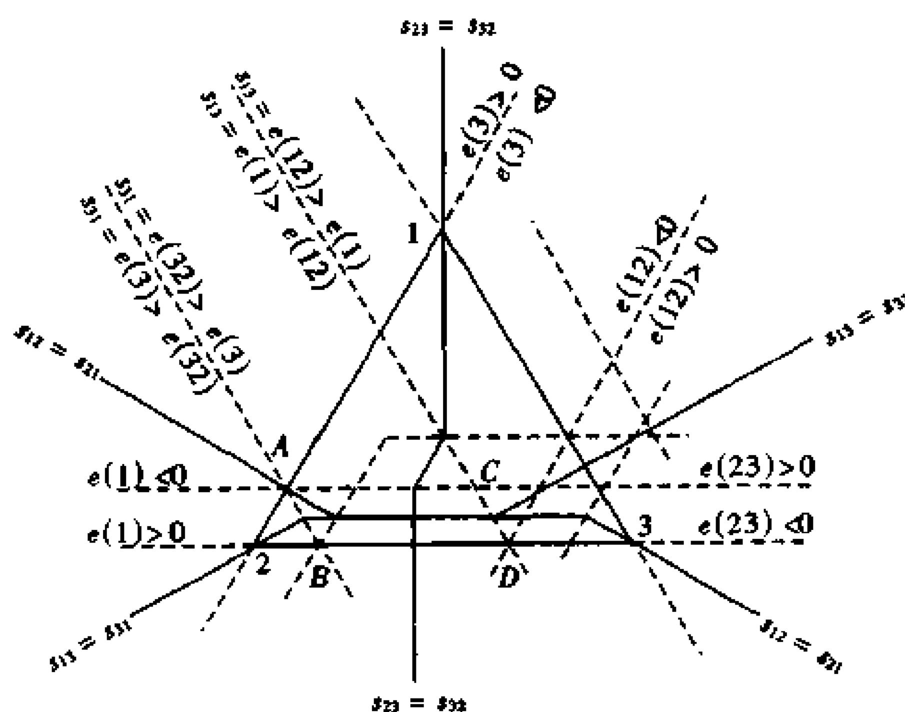


图 4-5

的图形，先考虑 $s_{13}(x)$. 由定义可知，

$$s_{13}(x) = \max\{e(1), e(12)\}.$$

在图4-5中，以直线 CD 为界，直线的左边 $e(1) > e(12)$，所以 $s_{13} = e(1)$；直线的右边 $e(12) > e(1)$，所以 $s_{13}(x) = e(12)$.

同理，在直线 AB 的左边和右边分别有 $s_{31}(x) = e(3)$ 和 $s_{31}(x) = e(32)$.

因此，广义转归集 $X^*(\Gamma)$ 可以分成三个区域. 对于 AB 左边的区域，有

$$s_{13} = e(1), \qquad s_{31} = e(3).$$

因而 $s_{13} = s_{31}$ 的图形就是直线 $e(1) = 0$ 和 $e(3) = 0$ 的交角的等分角线. 其次，在直线 AB 和 CD 之间的区域中，

$$s_{13} = e(1), \qquad s_{31} = e(32).$$

因而 $s_{13} = s_{31}$ 的图形是距直线 $e(1) = 0$ 和 $e(32) = 0$ 等远的直线段. 同理，在 CD 右边的区域中，$s_{13} = s_{31}$ 的图形是直线 $e(12) = 0$ 和 $e(32) = 0$ 的交角的等分角线. 这样，就得到了 $s_{13}(x) = s_{31}(x)$ 的整个图形，它是由三条直线段组合起来的折线.

按照同样的方法，可以画出

$$s_{12}(x) = s_{21}(x) \tag{4-17}$$

和

$$s_{23}(x) = s_{32}(x) \tag{4-18}$$

的图形，它们都是一些折线. (4-16) 式、(4-17) 式、(4-18) 式的图形交于一点. 这就是说，在这一点处，(4-10) 式对于每一对局中人 i 和 j 成立. 根据定义 19 和定义 20，每两个局中人 i 和 j 在上述交点处都平衡. 因此，这个交点既属于对策的预核 $K^*(\Gamma)$，又属于它的核 $K(\Gamma)$，并且其他点都不属于二者. 不难算出，

$$K^*(\Gamma) = K(\Gamma) = \left\{x \mid x_1 = \frac{1}{12}, x_2 = \frac{13}{24}, x_3 = \frac{9}{24}\right\}.$$

这一点是对策的最小核心 $LC(\Gamma) = C_{-1/12}(\Gamma)$ 的中点，也是核心 $C(\Gamma)$ 的几何中心(参看例 3).

例 6 例 4 中三人合作对策 Γ 的特征函数 v 的值为

$$v(i) = 0 \qquad (i = 1,2,3),$$
$$v(12) = 4, \qquad v(13) = 3,$$
$$v(23) = 10, \qquad v(123) = 8.$$

前已指出，$C(\Gamma) = \varnothing$. 令 ε 从 0 逐渐增大，当 $\varepsilon = 1$ 时，强 ε 核心开始出现，它位于转归三角形 $X(\Gamma)$ 的外面，这就是最小核心 $LC(\Gamma) = C_1(\Gamma)$，如图 4-6 所示.

采用与例 5 相同的方法，画出

$$s_{12}(x) = s_{21}(x), \tag{4-19}$$

$$s_{13}(x) = s_{31}(x), \tag{4-20}$$

$$s_{23}(x) = s_{32}(x). \tag{4-21}$$

的图形. 这三个方程的图形交于一点. 这一点就是构成对策 Γ 的预核 $K^*(\Gamma)$ 的唯一广义转归.

在核 $K(\Gamma)$ 处，只有(4-21)式一个等式成立，另外两个换成了不等式：

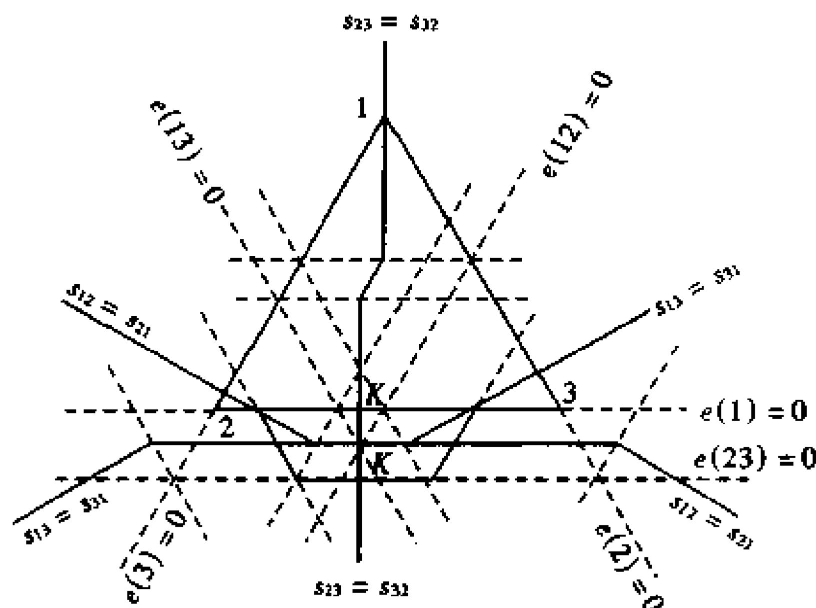


图 4-6

$$s_{21}(x) > s_{12}(x), \qquad x_1 = 0 = v(1),$$
$$s_{31}(x) > s_{13}(x), \qquad x_1 = 0.$$

这两个不等式正是(4-14) 式或(4-15) 式的形式.

不难算出,

$$K^*(\Gamma) = \{x \mid x_1 = -1, x_2 = 5, x_3 = 4\}.$$

同例 5 的情形一样,它也是最小核心 $LC(\Gamma) = C_1(\Gamma)$ 的中点.

对策的核为

$$K(\Gamma) = \{x \mid x_1 = 0, x_2 = 4.5, x_3 = 3.5\}.$$

4.9 核　　仁

定义 21　设 $\Gamma = [I, v]$ 是 n 人合作对策. 对于每一个 $x \in X(\Gamma)$, 定义一个 2^n 维向量如下:

$$\theta(x) = (\theta_1(x), \theta_2(x), \cdots, \theta_{2^n}(x)),$$

其中

$$\theta_i(x) = e(S, x) \qquad (S \subset I),$$
$$\theta_1(x) \geqslant \theta_2(x) \geqslant \cdots \geqslant \theta_{2^n}(x).$$

定义 22　设 $x, y \in X(\Gamma)$. 若存在下标 k_0, 使得

$$\theta_k(x) = \theta_k(y) \qquad (k < k_0),$$
$$\theta_{k_0}(x) < \theta_{k_0}(y),$$

则称 $\theta(x)$ 的字典次序在 $\theta(y)$ 之前，或者说，$\theta(x)$ 在字典次序上小于 $\theta(y)$，记为

$$\theta(x) \prec \theta(y).$$

"非 $\theta(y) \prec \theta(x)$" 则记为 $\theta(x) \preceq \theta(y)$.

定义 23 n 人合作对策 $\Gamma = [I, v]$ 的核仁是转归 x 的集 $N(\Gamma)$，在其中的每一个 x 处，θ 在字典次序上为最小，即

$$N(\Gamma) = \{x \mid x \in X(\Gamma); 对于一切\ y \in X(\Gamma), \theta(x) \preceq \theta(y)\}.$$

核仁的概念是由施迈德勒(D. Schmeidler) 首先提出来的.

定理 8 n 人合作对策 $\Gamma = [I, v]$ 的核仁 $N(\Gamma)$ 非空.

定理 9 设 n 人合作对策 $\Gamma = [I, v]$ 的核仁是 $N(\Gamma)$，则 $|N(\Gamma)| = 1$，即 $N(\Gamma)$ 由唯一的一个转归构成.

定理 10 设 n 人合作对策 $\Gamma = [I, v]$ 的核是 $K(\Gamma)$，核仁是 $N(\Gamma)$，则 $N(\Gamma) \subset K(\Gamma)$.

根据定理 10，如果一个 n 人合作对策 Γ 的核 $K(\Gamma)$ 只包含一个点，则这个点也就是 Γ 的核仁. 例如上述例 5 和例 6 中对策的核都只含有一个点，因而对策的核仁也就是这个点.

定理 10 不仅说明了 n 人合作对策的核仁 $N(\Gamma)$ 含在核 $K(\Gamma)$ 之中，而且为 $K(\Gamma)$ 的存在性提供了一种证明，比其他直接的证明方法简单得多.

下面介绍与核仁等价的一个概念.

定义 24 设 $\Gamma = [I, v]$ 是 n 人合作对策. 记

$$X^0 = X(\Gamma), \quad \Sigma^0 = \{S \mid S \subset I, S \neq \varnothing, I\}.$$

按照下述方式构造转归集的序列

$$X^0 \supset X^1 \supset \cdots \supset X^\kappa$$

和联盟集的序列

$$\Sigma^0 \supset \Sigma^1 \supset \cdots \supset \Sigma^\kappa.$$

对于 $k = 1, 2, \cdots, \kappa$，定义

$$\epsilon^k = \min_{x \in X^{k-1}} \max_{S \in \Sigma^{k-1}} e(S, x), \tag{4-22}$$

$$X^k = \{x \mid x \in X^{k-1}, \max_{S \in \Sigma^{k-1}} e(S, x) = \epsilon^k\}, \tag{4-23}$$

$$\Sigma_k = \{S \mid S \in \Sigma^{k-1}, e(S, x) = \epsilon^k, \forall x \in X^k\}, \tag{4-24}$$

$$\Sigma^k = \Sigma^{k-1} \backslash \Sigma_k, \tag{4-25}$$

其中 κ 是使得 $\Sigma^k = \varnothing$ 的第一个 k 值.

集 X^κ 称为 Γ 的字典中心.

定理 11 定义 24 中的 κ 为有限数. 对于 $k = 1, 2, \cdots, \kappa$，有

1° ϵ^k 为有限数；

2° X^k 为非空紧凸集；

3° $\Sigma_k \neq \varnothing$；

而对于 $k = 1, 2, \cdots, \kappa - 1$，有

4° $\varepsilon^{k+1} < \varepsilon^k$.

定理 12　n 人合作对策 $\Gamma = [I, v]$ 的字典中心由唯一的一个转归构成.

定理 13　n 人合作对策 $\Gamma = [I, v]$ 的核仁等于它的字典中心.

n 人合作对策 $\Gamma = [I, v]$ 的字典中心的构造为核仁的计算提供了一种方法.其算法可以通过解一系列线性规划来实现:

首先,对于 $k = 1$,考虑(4-22)式至(4-24)式.由(4-22)式有

$$\varepsilon^1 = \min_{x \in X^0} \max_{S \in \Sigma^0} e(S, x).$$

令

$$\max_{S \in \Sigma^0} e(S, x) = \alpha,$$

则

$$\alpha \geqslant e(S, x) = v(S) - x(S) \quad (S \in \Sigma^0),$$

即

$$x(S) + \alpha \geqslant v(S) \quad (S \in \Sigma^0).$$

确定 ε^1, X^1 和 Σ_1 的问题等价于解下列线性规划问题:

$$\left.\begin{aligned} &\min \alpha \\ &\text{s.t.} \quad x(S) + \alpha \geqslant v(S) \quad (S \in \Sigma^0), \\ &\qquad\quad x \in X^0. \end{aligned}\right\} \tag{4-26}$$

这一线性规划问题的极小值是 ε^1,它是极大超出值的极小值,在一个转归集 X^1 上,通过联盟集 Σ_1 中的联盟达到这个值 ε^1.这就是说,对于一切 $S \in \Sigma_1$ 和一切 $x \in X^1$,有

$$e(S, x) = \varepsilon^1.$$

其次,撤开 Σ_1 中的联盟,考虑下列线性规划问题:

$$\begin{cases} \min \alpha \\ \text{s.t.} \quad x(S) + \alpha \geqslant v(S) \qquad (S \in \Sigma^1 = \Sigma^0 \setminus \Sigma_1), \\ \qquad\quad x \in X^1. \end{cases} \tag{4-27}$$

这一线性规划将给出第二大的超出值 ε^2,以及相应的(4-23)式的 X^2 和(4-24)式的 Σ_2.再撤开 Σ_2 中的联盟,重复以上步骤,直到不再有联盟剩下为止.最后得到唯一的一个转归,它就是对策 Γ 的核仁中的唯一元素.

例 7　考虑例 5 中的三人合作对策 Γ.特征函数 v 的值为

$$v(i) = 0 \qquad (i = 1, 2, 3),$$

$$v(12) = \frac{1}{3}, \qquad v(13) = \frac{1}{6},$$

$$v(23) = \frac{5}{6}, \qquad v(123) = 1.$$

由第一次线性规划(4-26)式,给出

$$\alpha = \varepsilon^1 = -\frac{1}{12},$$

$$X^1=\{x \mid x_1=\frac{1}{12}, x_2 \leqslant \frac{9}{12}, x_3 \leqslant \frac{7}{12}, x_2+x_3=\frac{11}{12}\},$$
$$\Sigma_1=\{\{1\},\{2,3\}\},$$
$$\Sigma^1=\Sigma^0 \setminus \Sigma_1=\{\{2\},\{3\},\{1,2\},\{1,3\}\}.$$
集 X^1 就是 Γ 的最小核心.

由第二次线性规划(4-27)式得到

$$\alpha=\epsilon^2=-\frac{7}{24},$$
$$X^2=\{x \mid x_1=\frac{1}{12}, x_2=\frac{13}{24}, x_3=\frac{9}{24}\},$$
$$\Sigma_2=\{\{1,2\},\{1,3\}\},$$
$$\Sigma^2=\Sigma^1 \setminus \Sigma_2=\{\{2\},\{3\}\}.$$

类似地,由第三次线性规划,可得到

$$\alpha=\epsilon^3=-\frac{9}{24},$$
$$X^3=X^2,$$
$$\Sigma_3=\{\{3\}\},$$
$$\Sigma^3=\Sigma^2 \setminus \Sigma_3=\{\{2\}\}.$$
由第四次也是最后一次线性规划,给出

$$\alpha=\epsilon^4=-\frac{13}{24},$$
$$X^4=X^3=X^2,$$
$$\Sigma_4=\{\{2\}\},$$
$$\Sigma^4=\Sigma^3 \setminus \Sigma_4=\varnothing.$$

因此,对策的核仁为

$$N(\Gamma)=X^4=\{(\frac{1}{12},\frac{13}{24},\frac{9}{24})\}.$$

例 8　设四人合作对策 Γ 的特征函数 v 的值为

$$v(i)=0 \qquad (i=1,2,3,4),$$
$$v(12)=v(34)=v(14)=v(23)=1,$$
$$v(13)=\frac{1}{2}, \quad v(24)=0,$$
$$v(123)=v(124)=v(134)=v(234)=1,$$
$$v(1234)=2.$$

通过解三个线性规划问题,就可以求出对策的核仁.容易验证,从

$$X^0=X(\Gamma)$$
和

$$\Sigma^0=\{S \mid S \subset I, S \neq \varnothing, I\}$$
出发,有

1°
$$\varepsilon^1 = 0,$$
$$X^1 = \{(1-\frac{3}{4}\lambda, \frac{3}{4}\lambda, 1-\frac{3}{4}\lambda, \frac{3}{4}\lambda)\} \quad (0 \leqslant \lambda \leqslant 1),$$
$$\Sigma_1 = \{\{1,2\},\{3,4\},\{1,4\},\{2,3\}\},$$
$$\Sigma^1 = \Sigma^0 \setminus \Sigma_1 = \{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\},$$
$$\{1,3\},\{2,4\},\{1\},\{2\},\{3\},\{4\}\}.$$

2°
$$\varepsilon^2 = -\frac{1}{2},$$
$$X^2 = \{(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})\},$$
$$\Sigma_2 = \{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\},\{1,3\},\{1\},\{2\},\{3\},\{4\}\},$$
$$\Sigma^2 = \Sigma^1 \setminus \Sigma_2 = \{\{2,4\}\}.$$

3°
$$\varepsilon^3 = -1,$$
$$X^3 = X^2 = \{(1/2, 1/2, 1/2, 1/2)\},$$
$$\Sigma_3 = \{\{2,4\}\},$$
$$\Sigma^3 = \Sigma^2 \setminus \Sigma_3 = \varnothing.$$

因此,
$$N(\Gamma) = X^3 = \{(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})\}.$$

4.10 沙普利值

沙普利(L.S.Shapley)于1952年发现,对于一个 n 人合作对策 $\Gamma=[I,v]$,从三条公理出发,可以确定唯一的一组值,这组值可以作为分配给 Γ 的全体局中人的值.

定理 14(沙普利定理) 设 $\Gamma=[I,v]$ 是 n 人合作对策,存在唯一的一组沙普利值:
$$\Phi_i(v) = \sum_{S \ni i} \frac{(n-|S|)!(|S|-1)!}{n!}[v(S) - v(S \setminus i)] \quad (i = 1,2,\cdots,n).$$

例 9 例7中三人合作对策 Γ 的特征函数 v 的值为
$$v(i) = 0 \qquad (i = 1,2,3),$$
$$v(12) = \frac{1}{3}, \qquad v(13) = \frac{1}{6},$$
$$v(23) = \frac{5}{6}, \qquad v(123) = 1.$$

这个对策的沙普利值为
$$\Phi(v) = \left(\frac{5}{36}, \frac{17}{36}, \frac{14}{36}\right).$$

例 10 对策论文献中常常引用的一个例子是所谓"投票对策".五个局中人,局中人1有三张投票权,其余的局中人2,3,4,5各有一张投票权.自由结成联盟

后，总票数过半即可获胜.

如果把这个合作对策 Γ 用(0,1)规范化特征函数 v 表示出来，以 $v=1$ 表示获胜，$v=0$ 表示失败，则

$$v(12)=v(13)=v(14)=v(15)=1,$$
$$v(123)=v(124)=v(125)=v(134)=v(135)=v(145)=1,$$
$$v(1234)=v(1235)=v(1245)=v(1345)=v(2345)=1,$$
$$v(12345)=1,$$
$$v(S)=0 \quad (\text{对于其他的 } S\subset I).$$

容易算出，沙普利值为

$$\Phi(v)=\left(\frac{6}{10},\frac{1}{10},\frac{1}{10},\frac{1}{10},\frac{1}{10}\right).$$

这个对策的核和核仁为

$$K(\Gamma)=N(\Gamma)=\left\{\left(\frac{3}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7}\right)\right\}.$$

参考文献

1 王建华著.对策论.北京:清华大学出版社,1986.

2 Owen G. Game theory. 2nd ed. New York: Academic Press, 1982.

3 Rosenmüller J. The theory of games and markets. Amsterdam: North-Holland Publishing Co, 1981.

3 Szép J, Forgó F. Introduction to the theory of games. Holland: D Reidel Publishing Co, 1985.

4 Thomas L C. Games, theory and applications. Chichester: Ellis Horwood, 1984.

5 Vorob'ev NN. Game theory, lectures for economists and systems Scientists. New York: Springer-Verlag, 1977.

6 Wang Jianhua. The theory of Games. Oxford mathematical monographs. Oxford: Oxford University Press, 1988.

·经济数学卷·

第18篇

信息论

编　者　孟庆生
审校者　肖国镇

目　录

引　言

近代信息论，也称狭义信息论，是自1948年美国工程师申农（C. E. Shannon）的开创性论文[①]发表以后发展起来的，至今已有半个世纪的历史．这期间它经历了理论的确立、发展与逐步完善的过程，目前已被广泛地应用于通信技术以及其他领域．其主要内容包括信息理论、编译码技术及密码学三部分．

信息是反映世界（包括自然界和人类社会）诸种事物多样性的一种概念．人们在认识世界和改造世界中，在社会活动的交往中，都在不断地获取信息并传递信息；然而"信息"的本质特征是什么呢？又如何作出合适的信息度量呢？信息有着不同的形态，诸如语声、文字、符号及各种图像等，它的本质是随机的；也就是说，信息蕴含于不肯定性中．简而言之，**随机变量**就是信息；普通的一张报纸或电视上一帧画面，都只是某类信息（随机变量）的一个"样本"．在这种观念下，"**信息量**"可被确定为相应随机变量的"**数学期望**"。

申农的信息量，不考虑事件发生的时间、地点、内容以及人们对该事件的态度和反应，而只顾及事件发生的状态数目和每种状态发生的可能性大小，这就使信息度量具有普遍意义及相当广泛的适用性．

信息有了定量表征，就可进行存储和传递．信息论研究的基本问题是，如何有效且可靠地存储和传递信息，并提供实现的方法和技术．

1　信　息　量

1.1　熵

1.1.1　基本概念

定义1　设 X 是取有限个值的随机变量，分布律为

$$P\{X=x_i\}=p_i \qquad (i=1,2,\cdots,n),$$

则 X 的**申农熵**为

$$H(X)\xlongequal{\text{def}}-\sum_{i=1}^{n}p_i\log_a p_i, \tag{1-1}$$

---

① Shannon C E. A Mathematical theory of communication. Bell system Tech, 1948(27): 379～423, 623～656

其中对数的底 a 可取任意正整数，并规定当 $p_i = 0$ 时，$-p_i\log_a p_i = 0$.

底数 a 决定了熵的单位. 特别地，对于 $a = 2, e, 10$，熵的单位分别称为比特、奈特、哈特利；由换底公式

$$\log_a x = \frac{\log_b x}{\log_b a},$$

可得，上述各单位之间的换算关系为

$$1\quad \text{Hartley} \approx 3.32\ \text{bit},$$
$$1\quad \text{nat} \approx 1.44\ \text{bit}.$$

下面如无特别声明，均取 $a = 2$，且记 $\log_2 x = \text{lb}x$；当 $a = e$ 时，简记 $\log_e x = \ln x$.

例1 设数字电视屏上有 $500 \times 600 = 3 \times 10^5$ 个格点，按每点有10个不同的亮度等级计算，则共可组成 $10^{3\times10^5}$ 个不同的画面；依等概率计算，平均每个画面提供的信息量（申农熵）为

$$H_1 = -\text{lb}10^{-3\times10^5} = 3 \times 10^5 \times 3.32\text{bit} \approx 10^6\text{bit}.$$

另外，假定一篇千字文章中的每个字，可从万字表中任意选择，则能组成不同的千字文章数目为

$$N = 10000^{1000}\text{篇} = 10^{4000}\text{篇},$$

也按等概率计算，平均每篇千字文章可提供的信息量为

$$H_2 = \text{lb}N = \text{lb}10^{4000} = 4 \times 10^3 \times 3.32\text{bit} \approx 1.3 \times 10^4\text{bit}.$$

可见，“一个电视画面”提供的信息量，远远超过“一篇千字文”.

1.1.2 熵的性质

(1) 基本不等式.

1° 对于任意实数 $x > 0$，有

$$1 - \frac{1}{x} \leqslant \ln x \leqslant x - 1, \tag{1-2}$$

其中等号成立，当且仅当 $x = 1$；

2° 对于 $p_i \geqslant 0, q_i \geqslant 0, \sum_{i=1}^{n} p_i = 1 = \sum_{i=1}^{n} q_i$，有

$$-\sum_{i=1}^{n} p_i \ln p_i \leqslant -\sum_{i=1}^{n} p_i \ln q_i, \tag{1-3}$$

或等价地，可写为

$$\sum_{i=1}^{n} p_i \ln\frac{p_i}{q_i} \geqslant 0,$$

其中等号成立，当且仅当 $p_i = q_i, i = 1, 2, \cdots, n$；

3° 对于 $u_i > 0, v_i > 0$，有

$$\sum_{i=1}^{n} u_i \ln \frac{u_i}{v_i} \geqslant \left(\sum_{i=1}^{n} u_i\right) \ln \frac{\left(\sum_{i=1}^{n} u_i\right)}{\left(\sum_{i=1}^{n} v_i\right)} \tag{1-4}$$

其中等号成立,当且仅当

$$\frac{u_i}{v_i} = \left(\sum_{j=1}^{n} u_j\right) \Big/ \left(\sum_{j=1}^{n} v_j\right) \quad (i = 1,2,\cdots,n).$$

(1-2) 式可由函数 $f(x) = x - 1 - \ln x$ 的导数来判定;(1-3)、(1-4) 式则可依(1-2) 式相继获得;下述熵的基本性质均可由申农熵定义式(1-1) 及以上三个基本不等式直接获得.

(2) $$H(X) \equiv H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i \geqslant 0,$$

其中等号成立,当且仅当存在 k,使 $p_k = 1$,其他的 $p_i = 0, i \neq k$.

这表明,确定的概率场(无随机性) 的熵最小,也就是说,对于完全确定性的现象,认为没有提供有效的信息量.

特别需要指出,申农熵只与随机变量的分布律$(p_1, p_2, \cdots, p_n)$ 有关,而与该随机变量的取值无关,故也记为 $H(X) \equiv H(p_1, p_2, \cdots, p_n)$,实质上 $H(X)$ 是个 n 元变量 $p_1, p_2, \cdots, p_n$ 的非负函数.

(3) $$H(p_1, p_2, \cdots, p_n) \leqslant \ln n,$$

其中等号成立,当且仅当 $p_i = \frac{1}{n}, i = 1,2,\cdots,n$. 这表明,等概率场具有最大熵.

(4) 设随机变量 X 的分布律为

$$X:\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},$$

另外,$Y = f(X)$ 作为 X 的函数(多一映射),决定基本事件$\{x_1, x_2, \cdots, x_n\}$ 的一个分类:

$$A_1, A_2, \cdots, A_m, \quad A_k \cap A_l = \varnothing \quad (k \neq l),$$

其中 x_i 与 x_j 同属于一个 A_k,当且仅当

$$f(x_i) = f(x_j) \equiv y_k.$$

于是随机变量 Y(不一定取实数值) 可表示为

$$Y:\begin{pmatrix} y_1 & y_2 & \cdots & y_m \\ q_1 & q_2 & \cdots & q_m \end{pmatrix},$$

其中 $q_k \xlongequal{\text{def}} P\{Y = y_k\} = \sum_{x_i \in A_k} p_i \quad (k = 1,2,\cdots,m)$.

随机变量函数的熵不会增加,即 $H(f(X)) \leqslant H(X)$. 这表明,随机场分辨率越低,其信息量越少.

(5) 可加性.

$H(p_{11}, p_{12}, \cdots, p_{1k_1}, p_{21}, p_{22}, \cdots, p_{2k_2}, \cdots, p_{n1}, \cdots, p_{nk_n})$

$$= -\sum_{i=1}^{n}\sum_{k=1}^{k_i} p_{ik}\ln p_{ik}$$

$$= H(q_1, q_2, \cdots, q_n) + \sum_{i=1}^{n} q_i H\left(\frac{p_{i1}}{q_i}, \cdots, \frac{p_{ik_i}}{q_i}\right), \quad (1\text{-}5)$$

其中 $p_{ik} \geqslant 0, q_i = \sum_{k=1}^{k_i} p_{ik} > 0, \sum_{i=1}^{n} q_i = 1.$

(1-5) 式的信息含义是,随机变量 X 的熵 $H(p_{11}, p_{12}, \cdots, p_{1k_1}, p_{21}, p_{22}, \cdots, p_{2k_2}, \cdots, p_{n1}, \cdots p_{nk_n})$,等于 X 的函数 $Y = f(X)$ 的熵 $H(q_1, q_2, \cdots, q_n)$ 再加上已知 Y 时 X 的"条件熵".另外,在一定意义下,熵的可加性决定了其表达式的唯一性.

下面对条件熵给出明确表达,同时叙述另一个重要的信息量"互信息"的概念.

1.2 互 信 息

1.2.1 条件熵

设二随机变量 X 与 Y 均取有限个值,联合分布为

$$p(x, y) = P\{X = x, Y = y\} \quad (x \in \mathscr{X}, y \in \mathscr{Y}),$$

其中 $\mathscr{X}$ 与 $\mathscr{Y}$ 是含有限个元素的非空集合.

定义 2 若已知 $\{Y = y\}$,则

$$H(X/Y = y) \xlongequal{\text{def}} \sum_{x \in \chi} p(x/y) \log_a \frac{1}{p(x/y)} \quad (1\text{-}6)$$

称为随机变量 X 的**条件熵**,其中

$p(x/y) = p(x, y)/p(y)$ 为条件概率,假定 $p(y) = P\{Y = y\} > 0.$

定义 3 若已知随机变量 Y 下,则

$$H(X/Y) \xlongequal{\text{def}} \sum_{y \in \mathscr{Y}} p(y) H(X/Y = y) = \sum_{x, y} p(x, y) \log_a \frac{1}{p(x/y)} \quad (1\text{-}7)$$

称为 X 的**平均条件熵**,其中求和遍及 X 与 Y 的值域 $\mathscr{X}$ 与 $\mathscr{Y}$.以后简称 $H(X/Y)$ 为条件熵.

条件熵有如下性质:

1° 法诺不等式

设 X 与 Y 是具有同样值域 $\mathscr{X}$ 的二随机变量,令 $p_e = P\{X \neq Y\}$,则

$$H(X/Y) \leqslant H(p_e) + p_e \log_a(|\mathscr{X}| - 1) \quad (1\text{-}8)$$

称为法诺(Fano)不等式.其中 $|\mathscr{X}|$ 表示集合 $\mathscr{X}$ 中的元素个数;$H(x)$ 为熵函数,规定为

$$H(x) \xlongequal{\text{def}} -x\log_a x - (1 - x)\log_a(1 - x) \quad (0 \leqslant x \leqslant 1). \quad (1\text{-}9)$$

法诺不等式在逆编码定理中有用,其信息含义是:设想 X 为发送的消息,而 Y

为接收的消息，则条件熵 $H(X/Y)$ 表示收到 Y 时 X 的不肯定性（消息传输中的信息损失）．法诺不等式(1-8)表明，这种信息损失不超过这样两种信息量：其一，是否产生差错之不定度；其二，在已知发生误差的情况下，到底哪一个是真传信号之不定度．

2° $H(X/Y) \leqslant H(X)$， (1-10)

这表示，已知 Y 下，X 的不定度，绝不会超过 X 的原始不定度，纵然 Y 没能提供关于 X 的任何信息．

1.2.2 互信息

由(1-10)式可见，在已知 Y 条件下，X 的不定度 $H(X/Y)$ 小于 X 的原始不定度 $H(X)$，这表示 Y 中包含了 X 的某些信息，而且这个差额具有对称性．即

$$\begin{aligned} H(X) - H(X/Y) &= H(Y) - H(Y/X) \\ &= \sum_{x,y} p(x,y)\log_a \frac{p(x,y)}{p(x)p(y)}, \end{aligned} \quad (1\text{-}11)$$

其中 $p(x)$ 及 $p(y)$ 分别为随机变量 X 及 Y 的分布概率，即

$$\begin{aligned} p_x(x) &= P[X = x] \equiv p(x), \\ p_Y(x) &= P[Y = y] \equiv p(y). \end{aligned} \quad (1\text{-}12)$$

以后在不致混淆时，均用 $p(x)$，$p(y)$ 表示相应的分布概率．另外，对数的底 a 总取正数，且 $a > 1$，并简记为

$$\log_a x \equiv \log x. \quad (1\text{-}13)$$

在诸多信息量之关系式中均取同一对数的底．

定义 4 据(1-11)式，Y 中包含 X 的信息量与 X 中包含 Y 的信息量相同，即

$$I(X;Y) == \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)}, \quad (1\text{-}14)$$

称之为**互信息**，其中

$$p(x,y) = P\{X = x, Y = y\}$$

为 X 与 Y 的联合分布概率；

$$H(X,Y) \xlongequal{\text{def}} \sum_{x,y} p(x,y)\log \frac{1}{p(x,y)}, \quad (1\text{-}15)$$

称为**联合熵**．

需指出，联合熵并没引进新概念，因为当初定义一个随机变量 X 的熵时，(1-1)式并没有排除其为向量的情况，且也未限定它一定取实数值．因而这里所谓联合熵的(1-15)式，只表明一种记号．

至此，定义了熵、条件熵及互信息三种信息量，它们都适用于只取有限个值的多维随机变量的情况，相互间有如下基本关系式：

(1) $I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X) \geqslant 0$，

其中等号成立，当且仅当 X 与 Y 独立；

(2) $H(X) + H(Y/X) = H(X,Y) = H(Y) + H(X/Y)$；

(3) $H(X,Y) = H(X) + H(Y) - I(X;Y)$．

需注意,条件熵及互信息均可用无条件熵来表示,就是说,只有无条件熵才是最基本的信息量.上述关系式类比集合论中的相应关系,如图 1-1 所示.

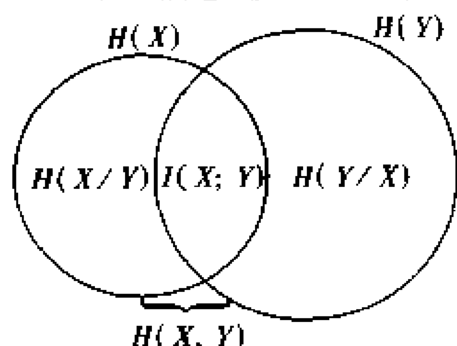


图 1-1

定理 1 (胡国定定理) 信息量的一个线性方程是恒等式,当且仅当相应的可加集函数的方程是恒等式①.

1.2.3 数据处理定理

定义 5 设有三个均取有限个值的随机变量 X, Y, Z,则

$$I(X, Y; Z) \overset{\text{def}}{=\!=} \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y, z)}{p(x, y) \cdot p(z)} \tag{1-16}$$

称为二随机变量 X 和 Y 与 Z 的互信息.

定义 6 称三个随机变量 X, Y, Z 构成马尔可夫链(Markov,简称**马氏链**),如果 $p(z/x, y) = p(z/y)$,对于一切使 $p(x, y, z) > 0$ 的 x, y, z 都成立.

马氏链的直观含义是,在已知"过去"和"现在"的条件下,"将来"只与"现在"有关.

定理 2 $I(X, Y; Z) \geqslant I(Y; Z)$,等号成立,当且仅当 X, Y, Z 构成马氏链.

定理 3 若 X, Y, Z 形成马氏链,则

$$I(X; Z) \leqslant I(X; Y), I(X; Z) \leqslant I(Y; Z).$$

定理 4 若四个随机变量 U, X, Y, V 形成马氏链,则

$$I(U; V) \leqslant I(X; Y). \tag{1-17}$$

此结果称为**数据处理定理**.其信息含义是,设想 $U—X—Y—V$ 构成通信系统,如图 1-2 所示.(1-17) 式表明,信息经过处理(编码和译码手续)之后,在无"边信息"的情况下(按马氏链考虑),信息有减无增.

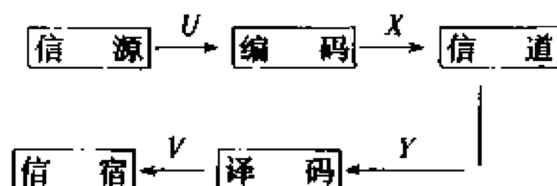


图 1-2

1.2.4 互信息的凸性

设二随机变量 X 与 Y 的联合分布概率为

$$p(x, y) = p(x)Q(y/x), \tag{1-18}$$

其中 $p(x) = P(X = x)$ 为**入口分布**(向量);$Q(y/x)$ 为已知 $\{X = x\}$ 下 $\{Y = y\}$ 的条件概率,并称之为**转移概率**(矩阵).这样一来,X 与 Y 的互信息(1-14) 式,就可写

① 胡国定. Теория Вероятностей и её применения. 概率论及其应用, T-7, 1-4, 1962, 441

为

$$I(X;Y)=\sum_{x,y}p(x)Q(y/x)\log\frac{Q(y/x)}{\sum_{x'}p(x')Q(y/x')};\tag{1-19}$$

可见,上式是多元变量$\{p(x)\}$及$Q(y/x)$的函数,于是可简记为

$$I(X;Y)=I(P;Q).\tag{1-20}$$

互信息$I(X;Y)$有如下凸性:

1° 互信息$I(X;Y)=I(P;Q)$是入口分布P的上凸函数($\cap$函数),即对于任意两个入口分布P_0及P_1,以及$0<\lambda<1$,均有

$$\lambda I(P_0;Q)+(1-\lambda)I(P_1;Q)\leqslant I(\lambda P_0+(1-\lambda)P_1;Q).\tag{1-21}$$

2° 互信息$I(X;Y)=I(P;Q)$是转移概率Q的下凸函数(U函数),即对任意二转移概率Q_0,Q_1及$0<\lambda<1$,均有

$$I(P;\lambda Q_0+(1-\lambda)Q_1)\leqslant\lambda I(P;Q_0)+(1-\lambda)I(P;Q_1).\tag{1-22}$$

需要指出,$\lambda P_0+(1-\lambda)P_0$也是一个入口分布;而$\lambda Q_0+(1-\lambda)Q_1$仍为转移概率.上述(1-21)式和(1-22)式均可用基本不等式证明.

1.3 关于信息量的几个问题

1.3.1 熵的唯一性

申农熵的定义(1-1)式,作为随机变量X的信息量或不肯定度$H(X)$,只是其分布律$(p_1,p_2,\cdots,p_n)$的一种多元函数,记为$H(X)=H(p_1,p_2,\cdots,p_n)$,而与$X$的取值并无关系,这在后面编码定理中还要进一步详述.现在问题是,还有没有其他更合适的函数形式作为这种信息量呢?对此,下述结论给出了否定的答案.

定理 5 (熵的唯一性) 假设

1° $H(p_1,p_2,\cdots,p_n)$是概率分布$(p_1,p_2,\cdots,p_n)$的多元连续函数;

2° $H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)=f(n)$是$n$的增函数;

3° 对于任意$p_{ij}\geqslant 0$,$p_i=\sum_{j=1}^{k_i}p_{ij}$,$\sum_{i=1}^{n}p_i=1$,

$$\begin{aligned}&H(p_{11},p_{12},\cdots,p_{1k_1},p_{21},\cdots,p_{2k_2},\cdots,p_{n1},\cdots,p_{nk_n})\\&=H(p_1,p_2,\cdots,p_n)+\sum_{i=1}^{n}p_iH\left(\frac{p_{i1}}{p_i},\frac{p_{i2}}{p_i},\cdots,\frac{p_{ik_i}}{p_i}\right),\end{aligned}\tag{1-23}$$

则有

$$H(p_1,p_2,\cdots,p_n)=C\sum_{i=1}^{n}p_i\log\frac{1}{p_i},\tag{1-24}$$

其中C是某常数.

在上述三条假设中,关键的是第3°条,它表示了“信息的可加性”;而假设2°的

信息含义是要求在等概率场时，不肯定度随着元数 n 的增加而变大，这在直觉上是自然的想法，三定一总要难于两择一；假设 1° 是要在连续函数中寻找熵的表达式，这也无可非议. 总之在这三条假设下，便可证明熵的表达(1-24) 式唯一给定，其中常数 C 虽然未定，但正好为信息量单位(bit, nat, Hartley) 的选择留有余地.

至此可以明确，以申农熵为基础的常用信息量共有四种：无条件熵、条件熵、互信息(这三种已分别由(1-1) 式、(1-7) 式及(1-14) 式给出)，以及**条件互信息**.

定义 7 若已知随机变量 Z，则

$$I(X;Y/Z) \xlongequal{\text{def}} \sum_{x,y,z} p(x,y,z)\log\frac{p(x,y,z)}{p(x/z)p(y/z)}, \tag{1-25}$$

称为 X 与 Y(Z 条件下) 的**条件互信息**. 它有下列性质：

1° $I(X;Y/Z) \geqslant 0$，等号成立，当且仅当 X, Z, Y 成马氏链；

2° $I(X;Y/Z) = H(X/Z) - H(X/Y,Z)$；

3° $I(X;Y,Z) = I(X;Z) + I(X;Y/Z)$.

关于有限随机场的信息量的确立及性质，文献中有许多研究，以上只是一些涉及申农熵的基本内容.

申农熵不只是在通信理论及若干交叉学科中有着广泛的应用，就是在一些古典的趣味问题中也显出了应用特色. 举例如下.

例 2 "假硬币"问题. 设有同一规格的硬币 25 个，其中 24 个是标准的，重量相同；另一个是假的，已知它比其他的稍轻一点，但外观却与标准币一样，难于辨认. 试问在不用砝码的天平上至少要称多少次，才能确认这个假硬币？

解 首先，该问题的不肯定性事件 E_0 是来自 25 个等可能事件中，其熵为 $H_0 = \log 25$. 每称量一次，只有三种结果：左偏、右偏或平衡，故所获之信息量

$$H_1 \leqslant \log 3;$$

若进行 k 次称量，则由复合试验结果 $A_k = E_1, E_2, \cdots, E_k$ 给出的关于事件 E_0 的互信息 $I(A_k;E_0)$ 为

$$\begin{aligned} I(A_k;E_0) &\leqslant H(A_k) = H(E_1,E_2,\cdots,E_k) \\ &\leqslant H(E_1) + H(E_2) + \cdots + H(E_k) = kH_1 \\ &\leqslant k\log 3. \end{aligned}$$

若 k 次试验正好确定了 E_0，则应有

$$H_0 = I(A_k;E_0),$$

即

$$\log 25 \leqslant k\log 3,$$

从而得

$$k \geqslant \frac{\log 25}{\log 3} = \log_3 25, \quad 3^k \geqslant 25,$$

取整数 $k \geqslant 3$；由此获知，至少要进行 3 次，才能确认假硬币.

事实上 3 次肯定可以称出假硬币来：为了首次称量能获得最大信息量，应使各组的结局尽可能有相近的概率，于是应将 25 个硬币分成 8,8,9 三组，假币在各组的概率分别为 8/25, 8/25 及 9/25，比较接近. 试验 E_1 是各选 8 个放在天平两端称重；接着 E_2 是在确认的 8 个(E_1 中偏轻一端) 或 9 个(E_1 中平衡时，取其余 9 个) 中再各

选3个放在天平两端上称重;最后试验 E_3 只是在3个或2个中定案了,这样3次肯定可认出假币来.

为建立信息量的直觉数码概念,下面举个简明例子.

例3 设基本事件空间为

$$\Omega = \{\omega = (\omega_1,\omega_2,\omega_3);\omega_i = 0,1\}.$$

概率为

$$P(\omega) = 1/8 \quad (\omega \in \Omega);$$

两个随机变量分别为

$$X(\omega) = \omega, Y(\omega) = \omega_1 \quad (\text{当 } \omega = (\omega_1,\omega_2,\omega_3)\text{ 时}),$$

求其信息量.

解 诸信息量为

$$H(X) = \mathrm{lb}8 = 3\mathrm{bit},$$
$$H(Y) = 1\mathrm{bit}, H(X/Y) = 2\mathrm{bit},$$
$$H(Y/X) = 0, I(X;Y) = 1\mathrm{bit}.$$

且可验证下列关系式:

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X).$$

1.3.2 申农熵的局限性

前述申农信息量只是对有限概率场,即只取有限个值的随机变量而定义的.试问能否将此定义推广到可列场,以及连续取值的随机变量的情况呢?简单回答是,对于可列场,申农熵有条件地存在;而对于取连续值的随机变量,基本上不存在上述含义的熵.

1. 可列场申农熵的存在性

定义8 设可列场概率分布为 $P = [p_1,p_2,\cdots,p_n,\cdots]$, $p_i \geqslant 0$, $\sum_{i=1}^{\infty} p_i = 1$,则

$$H(P) \xlongequal{\text{def}} -\sum_{i=1}^{\infty} p_i \log p_i, \tag{1-26}$$

称为**申农熵**.这里不同于有限场的是,(1-23)式不一定存在有限值.

例4 记 $A = \sum_{n\geqslant 2} \frac{1}{n(\log n)^s}$,若

$$p_n = \frac{1}{An(\log n)^s} \quad (n = 2,3,\cdots),$$

求其申农熵.

解 用级数收敛判别法可得

当 $1 < s \leqslant 2$ 时, $H(P) = +\infty$;

当 $s > 2$ 时, $H(P) < +\infty$.

一般地,有下述结论.

定理6 设 $P = [p_1,p_2,\cdots]$ 是概率分布, $p_n \geqslant 0$, $n = 1,2,\cdots$, $\sum_{n=1}^{\infty} p_n = 1$,则

1° 若$\sum_{n=1}^{\infty} p_n \log n < +\infty$,便有 $H(P) < +\infty$;

2° 反之,若 $H(P) < +\infty$,且$[p_n]$单调,即

$p_1 \geqslant p_2 \geqslant \cdots$,则有

$$\sum_{n=1}^{\infty} p_n \log n < +\infty. \tag{1-27}$$

这就是说,在分布律单调的条件下,(1-27)式是可列场申农熵存在的充要条件;若取消单调分布的条件,则虽然可列场的熵为有限,但(1-27)式也可以不成立.

2. 非离散型随机变量的申农熵问题

对于不是取有限值或可列值的随机变量 X 的情况,怎样定义申农熵呢?设 X 的分布函数为$F(x) = P\{X \leqslant x\}$,$T = [T_i, i = 1,2,\cdots]$是实数轴$\mathbf{R}^1$的一个分割,

$$T_i \cap T_j = \varnothing, \quad i \neq j, \quad \sum_{i=1}^{\infty} T_i = \mathbf{R}^1,$$

其中 T_i 皆为勒贝格(Lebesgue)可测集.记 X 的一个离散型量化"替身"为$[X]_T$,其概率分布为

$$P\{[X]_T = i\} = P\{X \in T_i\} = \int_{T_i} \mathrm{d}F(x) \equiv p_i,$$

于是$[X]_T$的熵

$$H([X]_T) \xlongequal{\text{def}} \sum_{i=1}^{\infty} p_i \log \frac{1}{p_i},$$

便与分割 T 有关.

定义 9 若 X 为上述随机变量,则

$$H(X) \xlongequal{\text{def}} \sup_T H([X]_T) \tag{1-28}$$

称为 X 的申农熵,其中上确界 sup 是对 $\mathbf{R}^1$ 的所有勒贝格分割 T 而取的.

上述定义可惜的是,连续型随机变量的申农熵全都不存在有限值.

定理 7 若随机变量 X 的分布函数 $F(x)$ 在实数集 R_1 上连续,则按(1-28)式定义的申农熵 $H(X) = +\infty$.①

1.3.3 广义信息量

上述结论表明,连续型随机变量不具有申农熵.那么,对于这类随机信号又如何给以信息表征呢?

1. 微分熵

定义 10 设连续型随机变量 X 的密度函数为$p(x)$,$x \in \mathbf{R}^1 = (-\infty, \infty)$,则

$$h(X) \xlongequal{\text{def}} -\int_{-\infty}^{\infty} p(x) \log p(x) \mathrm{d}x \tag{1-29}$$

---

① 孟庆生.关于 shannon 熵的局限性.工程数学学报,1986(2).

称为 X 的**微分熵**.

必须指出,如此之"微分熵"不是随机变量一一变换的不变量,因而不适宜作为"信息测度".一般难以想象,信号经过一一变换后会改变"信息量".申农熵不但是一一变换的不变量,而且还具有编码意义.

例 5 设随机变量 X 具有密度函数

$$p(x) = \frac{2}{\pi} \cdot \frac{1}{1+x^2} \quad (x \geqslant 0),$$

取 $Y = f(X) = e^{-X}$,其密度函数

$$q(y) = p(-\ln y)\frac{1}{y} = \frac{2}{\pi} \cdot \frac{1}{(1+\ln^2 y)y} \quad (0 < y \leqslant 1),$$

这里 X 与 Y 为一一变换关系,但经一一变换后,$0 < h(X) < +\infty$,而 $h(Y) = -\infty$,说明一一变换后改变了信息量.

一般说来,若 $Y = f(X)$ 为一一变换,且变换函数存在非零导数 $f'(x)$,则有

$$h(Y) = h(X) + \int_{-\infty}^{\infty} p(x)\log|f'(x)|\,dx, \tag{1-30}$$

其中 $p(x)$ 为连续型随机变量 X 的密度函数.由此可见,所谓微分熵不是一一变换的不变量之根本原因,就是密度函数是一个有量纲的量,在变换过程中可以产生尺度变形,致使量值有巨大差别(密度函数的量纲为 1/[长度单位]).

2. 最大熵原理

连续型随机变量的微分熵,虽然在本质上有别于有限值随机变量的申农熵,但仍具备一定的理论意义及应用价值.

定理 8 (最大熵原理),在一定约束下,选择概率函数$[p(x), x \in \mathscr{X}]$,使其对应的熵达极值:

$$\begin{cases} -\int_{\mathscr{X}} p(x)\log p(x)\,dx = \max \text{ 或 } \min; \\ \int_{\mathscr{X}} g_i(x)p(x)\,dx = a_i, \int_{\mathscr{X}} p(x)\,dx = 1 \quad i = 1,2,\cdots,m. \end{cases} \tag{1-31}$$

其中 a_i 为常数;$g_i(x)$ 为已知函数.另外,对于离散型随机变量的情况,上述积分改为求和,微分熵化为申农熵.

按变分法(对于离散情况,用拉普拉斯(Laplace)乘子法),令

$$L = -p\log p - \lambda_0 p - \sum_{i=1}^{m}\lambda_i g_i p,$$

则欧拉(Euler)方程(不妨取自然对数底)为

$$\frac{\partial L}{\partial p} = -\log p - 1 - \lambda_0 - \sum_{i=1}^{m}\lambda_i g_i = 0,$$

由此得概率函数为

$$p(x) = C \cdot \exp\left[-\sum_{i=1}^{m}\lambda_i g_i(x)\right], x \in \mathscr{X} \tag{1-32}$$

其中常数 $\lambda_1, \cdots, \lambda_m$;$C = \exp[-1-\lambda_0]$. C 可由约束条件确定如下:

$$\begin{cases} \int_x C \cdot \exp[-\sum_{i=1}^{m} \lambda_i g_i(x)] dx = 1; \\ \int_x C \cdot g_i(x) \cdot \exp\{-\sum_{i=1}^{m} \lambda_i g_i(x)\} dx = a_i, \quad i = 1,2,\cdots,m. \end{cases} \tag{1-33}$$

相应最大熵为

$$H_{\max} = 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i a_i. \tag{1-34}$$

需指出，方程(1-33)为超越方程，一般不易求解，需深入研究或进行参数估计. 但对于特殊问题也有明确结果：① 对于取值于有限区间$[a,b]$上的连续型随机变量，其均匀分布具有最大微分熵，② 在方差一定的条件下，对于取值于全数轴上的连续型变量，其正态分布具有最大微分熵.

3. 广义信息量

上述对于有限场的申农熵定义(1-1)式和推广到可列场的定义(1-26)式，以及关于连续型变量的微分熵定义(1-29)式，三者可统一写成如下形式，并称之为申农熵.

$$H(X) = E\log\frac{1}{P(X)} \overset{\text{def}}{=\!=} \begin{cases} \sum_{x\in\mathscr{X}} p(x)\log\frac{1}{p(x)}; \\ \int_x p(x)\log\frac{1}{p(x)}dx. \end{cases} \tag{1-35}$$

其中$[p(x), x\in\mathscr{X}]$为随机变量X的概率函数，它在X为离散型时，为分布概率；而在X为连续型时，则为概率密度函数. 这里$\mathscr{X}$是X的值域，也不必限制于一维实数，可以是n维欧氏空间.

一般说来，若$f(X)$是随机变量X的任一适当的函数，只要其数学期望$Ef(X)$存在，$Ef(X)$就可称之为X的一种信息量. 特别地，若取$f(X) = -\log p(X)$，$Ef(X)$就是申农熵；若取$f(X) = (X - EX)^2$，$Ef(X)$就是方差. 申农熵以外的信息量均称为**广义信息量**，而申农熵定义(1-1)式则独具编码意义.

2　信源编码

2.1　信源编码规则

2.1.1　信源的概念

一般说来，信源就是个概率场. 比如，全体汉字及其概率分布，26 个拉丁字母及其概率分布等，都是信源. 由于实际上各种文字及符号的使用都是一串字加上标点形成语句来表现的，因而把一列随机变量$U = \{U_1, U_2, \cdots\}$叫做**信源**，其中每个随机变量U_i, $i = 1,2,\cdots$，取值于某个集合$\mathscr{U} = \{u\}$中，并称之为**信源字母集**(或消

息集),它的元素个数用 $|\mathscr{U}|$ 表示,总设 $|\mathscr{U}| < \infty$.特别地,当 U_i 为独立同分布时,称之为**无记忆信源**;而当它们构成马氏链时,则称为**马氏信源**.另外,称有限长信源列,$U^k = (U_1, U_2, \cdots, U_k)$ 的一个样本 $u^k = (u_1, u_2, \cdots, u_k)$ 为 k 长消息,它是随机向量 U^k 的一个实现,k 长消息的全体个数为 $|\mathscr{U}|^k$.

2.1.2 码的概念

取自信源字母集 $\mathscr{U}$ 中的一串消息,通常是由文字、标点及各种代号等组成的,不适于存储与传递,因而总是用一种合适的符号代替消息元;代用符号的集合 $\mathscr{X} = \{x\}$ 称为信号集或**码符集**,且假定其中元素数目有限,即 $|\mathscr{X}| < \infty$.在数字化通信中,常选码符集 $\mathscr{X}$ 为某有限域,$\mathscr{X} = GF(q)$,即元素数目为 q 的伽罗瓦域(Galois Field).

k 长消息的一个码,就是一对映射 (f, φ),

$$f: \mathscr{U}^k \to \mathscr{X}^*, \quad \varphi: \mathscr{X}^* \to \mathscr{U}^k, \tag{2-1}$$

其中 $\mathscr{X}^*$ 是码符集 $\mathscr{X}$ 之元素的全体有限长序列的集合,且

$$\mathscr{X}^* = [x^n = (x_1, x_2, \cdots, x_n): x_i \in \mathscr{X}; \quad 1 \leqslant i \leqslant n, n = 1, 2, \cdots];$$

而 $\mathscr{U}^k$ 则为 k 长消息的集合,

$$\mathscr{U}^k = [u^k = (u_1, u_2, \cdots, u_k): u_i \in \mathscr{U}, \quad 1 \leqslant i \leqslant k].$$

每个 $f(u^k) = x$ 叫做一个**码字**,且称

$$f(\mathscr{U}^k) = [f(u^k): u^k \in \mathscr{U}^k]$$

为**码字集合**,有时也简称为**码**.将消息变为码字的映射 f 称为**信源翻码**,由码字转成消息的映射 φ 则称为**信源译码**,将 (f, φ) 合称为**编码**(也简称码).当 $\mathscr{X} = GF(2)$ 时,称 x 为**二元码**.若将 k 长消息翻成固定 n 长码字,又将 n 长码字译成 k 长消息,即

$$f: \mathscr{U}^k \to \mathscr{X}^n, \quad \varphi: \mathscr{X}^n \to \mathscr{U}^k,$$

则称 (f, φ) 为 k 到 n 的**分组码**,此时码字称为**定长码**(诸码字均等长度);否则称为**变长码**.

2.1.3 编码规则

由于编码是从消息变为码字,而后又从码字换成消息,因此,经过两次映射后,最终得到的消息与原来的消息就可能不尽一致,即可能有 u^k 不等于 $\varphi(f(u^k))$ 的情况,便产生了**编码误差**.码 (f, φ) 的**误差概率**定义为

$$e(f, \varphi) \xlongequal{\text{def}} P\{\varphi(f(U^k)) \neq U^k\}, \tag{2-2}$$

其中 $U^k = (U_1, \cdots, U_k)$ 表示 k **长随机消息**.经过编码后,保持消息不变的概率就是**保真度**,它与误差概率正好相反.

信源编码有如下保真度准则:

(1) 零误差　要求误差概率为零,即

$$e(f, \varphi) = 0 \quad 或 \quad P\{\varphi(f(U)) = U^k\} = 1.$$

(2) ε 概率误差　要求误差概率不超过一个小数 $0 < \varepsilon < 1$,即

$$e(f,\varphi) < \varepsilon \quad 或 \quad P[\varphi(f(U^k)) = U^k] \geqslant 1-\varepsilon.$$

(3) 平均距离误差 要求编码(f,φ)满足

$$\frac{1}{k}Ed(U^k,\varphi(f(U^k))) < \varepsilon \qquad (0<\varepsilon<1), \tag{2-3}$$

其中 $d(u_1^k,u_2^k)$ 为 k 长消息 u_1^k 与 u_2^k 之间的不同分量个数,称之为汉明(Hamming)距离.

这三个准则,按保真度要求是递弱的,即准则(1) 最强,而(3) 最弱.

2.1.4 码的分类

一般说来,信源翻码 f 是对于每列消息$(u_1,u_2,\cdots)$所对应的一列信号$(x_1,x_2,\cdots)$所形成的 $\mathscr{U}^\infty$ 到 $\mathscr{X}^\infty$ 的映射,即

$$f:\mathscr{U}^\infty \to \mathscr{X}^\infty;$$

反之,译码 φ 则是逆映射,即

$$\varphi:\mathscr{X}^\infty \to \mathscr{U}^\infty.$$

其中 $\mathscr{U}^\infty = \{(u_1,u_2,\cdots),u_i\in\mathscr{U}\}$ 为所有消息列;$\mathscr{X}^\infty = \{(x_1,x_2,\cdots),x_i\in\mathscr{X}\}$ 为全体信号列.严格说来,由(2-1) 式给出的那种映射(f,φ)称为 k 长消息的**分组码**,其中按码长相等与否又分成**定长分组码**与**变长分组码**.关于非分组码类,主要有**树码**和**卷积码**,它们同是具有树形结构的码,而卷积码则是由给定"生成元"形成的一种树码.

分组码类也有许多种,最重要的实用分组码类便是唯一可译码,其特点是,对于任意联接起来的一列码字,都能无二义地分解为原来的码字.对于这类码,总设 φ 是 f 的逆映射 $\varphi = f^{-1}$.

一个码字$(x_1,x_2,\cdots,x_n)$的词头,是指诸码符列$(x_1,x_2,\cdots,x_m)$,$1\leqslant m\leqslant n$;如果一个分组码的每一个码字都不是另一个码字的词头,就称该码为**前束码**.

一个唯一可译码,如果它的每个码字,不必等到后面的码字出现,都能按序即时译出,就称之为**即时码**.

定理 1 一个码为即时码的充要条件是它为前束码.

2.2 分组编码

2.2.1 定长编码定理

定理 2 设无记忆信源 $U=\{U_1,U_2,\cdots\}$ 具有公共分布 $p(u)$,$u\in\mathscr{U}$,令

$$n(k,\varepsilon) \xlongequal{\text{def}} \min[n:存在\ k\ 到\ n\ 二元分组码(f,\varphi),使\ e(f,\varphi)<\varepsilon],$$

则

$$\lim_{k\to\infty}\frac{n(k,\varepsilon)}{k} = H(U) \quad (\forall\varepsilon\in(0,1)), \tag{2-4}$$

称为定长编码定理,其中 $H(U)$ 是公共分布的熵,即

$$H(U) \xlongequal{\text{def}} -\sum_{u\in\mathscr{U}}p(u)\lg p(u). \tag{2-5}$$

注：无记忆信源具有这样一种"信息稳定性"，即当 k 足够大时，该信源大部分 k 长样本含有大致相同的"信息量"，而这个定量就是 $kH(U)$；如果用二元码来表示这种消息，便有码长 $n \approx kH(U)$，或 $n/k \approx H(U)$，也就是说，公共分布的熵 $H(U)$ 表示了每个消息字母平均所需二元码符号位数.这就揭示了申农熵的本质.

2.2.2 变长编码定理

定义 1 设变长分组翻码 $f:\mathscr{U}^k \to \mathscr{X}^*$，则

$$\bar{l}(f) \xlongequal{\text{def}} E\left[\frac{1}{k}l(f(U^k))\right] \tag{2-6}$$

称为变长分组翻码的平均码长，其中 $l(x)$ 为码字 $x \in \mathscr{X}^*$ 的长度，即码字 x 的分量个数.

定理3 设 $U=\{U_1,U_2,\cdots\}$ 是无记忆信源，公共分布为 $p(u)$，$u\in\mathscr{U}$，其熵为 $H(U)$，由(2-5)式表示，则任一唯一可译码 $f:\mathscr{U}^k \to \mathscr{X}^*$ 其平均码长 $\bar{l}(f)$ 具有下限，即

$$\bar{l}(f) \geqslant \frac{H(U)}{\log|\mathscr{X}|}, \tag{2-7}$$

并且一定存在前束码，满足

$$\bar{l}(f) < \frac{H(U)}{\lg|\mathscr{X}|} + \frac{1}{k}, \tag{2-8}$$

进一步，若分组码 $f:\mathscr{U}^k \to \mathscr{X}^*$；$\varphi:\mathscr{X}^* \to \mathscr{U}^k$ 满足条件

$$\frac{1}{k}E[d(U^k,\varphi(f(U^k)))] < \epsilon \leqslant \frac{1}{2}, \tag{2-9}$$

则有

$$\bar{l}(f) \geqslant \frac{H(U)}{\log|\mathscr{X}|} - \frac{1}{\lg|\mathscr{X}|}[\epsilon\log(|\mathscr{U}|-1) + H(\epsilon)] - \frac{|\log(dk)|^+}{k\log|\mathscr{X}|}, \tag{2-10}$$

其中

$$d = e\cdot\frac{\log|\mathscr{U}|}{\log|\mathscr{X}|};\ e = 2.7182818284\cdots;$$

$$H(\epsilon) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon);$$

$$|r|^+ \xlongequal{\text{def}} \max(r,0).$$

若 f 的值域唯一可译，则(2-10)式中最后一项可取消.

说明：定理3通称申农第一定理，或无干扰信源编码定理.对比定理2，(2-4)式表示二元定长分组码，在保真度准则(2)下，当消息长 $k\to\infty$ 时，平均码长 $\bar{l}(f)$ 收敛于 $H(U)$；而定理3指出，纵然用较弱的准则(3)，且允许用变长码，只要坚持用一个较小的平均误差，这一渐近界本质上不能改进了.另一方面，这种极限性能却可用前束码来达到，而且用零误差译码准则(1)，就是说，平均码长下界(2-7)式不能更好了，即使用较弱准则及变长码，但此界却可用好的前束码在强准则下达到.

2.2.3　逆编码定理

上述关于分组编码的两个定理，分别表明无记忆信源的公共熵 $H(U)$，本质上就是在一定误差限制条件下每个信源符号的平均码长. 现在进一步考虑 k 到 n 定长分组码(f,φ)，

$$f:\mathscr{U}^k \to \mathscr{X}^n,\quad \varphi:\mathscr{X}^n \to \mathscr{U}^k. \tag{2-11}$$

对于二元码，由(2-7)式有 $n/k \geqslant H(U)$，不过这是对唯一可译码而言的. 如果取消唯一可译的要求，能否编出平均长度更短的码字呢？回答是肯定的，但这样码的误差概率是没有保证的，一般有下述定理.

定理 4　设无记忆信源公共分布为 $p(x)$，$u \in \mathscr{U}$，熵为 $H(U)$，码符集为 $\mathscr{X}$，其中元素个数为 $q = |\mathscr{X}|$，对于(2-11)式的 k 到 n 定长分组码(f,φ)，记误差概率为

$$e(f,\varphi) = P[U^k \neq \varphi(f(U^k))], \tag{2-12}$$

1° 若对于某个 $\delta > 0$，有

$$n/k \geqslant (H(U) + \delta)/\log q, \tag{2-13}$$

则当 k 充分大时，$e(f,\varphi)$ 可任意小；

2° 反之，若对于任意 $\delta > 0$，有

$$n/k \leqslant (H(U) - \delta)/\log q, \tag{2-14}$$

则当 k 充分大时，误差概率 $e(f,\varphi)$ 可任意接近 1.

由定理 4 可知，对于二元定长分组码，平均每信源符号码长 n/k 以信源熵 $H(U)$ 为临界. 当 $n/k > H(U)$ 时，编码误差可任意小；当 $n/k < H(U)$ 时，必使误差概率增大到近于 1.

对于变长码，关于码长限制也有下列定理.

定理 5　存在一前束码 $f:\mathscr{U} \to \mathscr{X}^*$，其具有给定码字长度 $l(f(u)) = n(u)$，$u \in \mathscr{U}$ 的充要条件是下列不等式成立：

$$\sum_{u\in\mathscr{U}} |\mathscr{X}|^{-n(u)} \leqslant 1. \tag{2-15}$$

上不等式称为克拉夫特(Kraft)不等式，它直接给出了对前束码长的限定指标.

2.2.4　常用编码法

下面介绍三种典型的实用信源编码方法.

1. 哈夫曼编码法

这种方法是，先将信源符号分成接近等概率的 q 组，记为 $E_0, E_1, \cdots, E_{q-1}$. 为了编成 q 元码，将各组元素的首位码符分别记为$0,1,\cdots,q-1$；进一步，将上述一线组 E_i，$i = 0,1,\cdots,q-1$，分成 q 个概率大致相等的 q 个小组，记为 $E_{i0}, E_{i1}, \cdots, E_{i,q-1}$，再将此二线组中诸元的第二位码符分别记为 $0,1,\cdots,q-1$. 如此逐步分组编号，直到各信源符号均被编成 q 元码符为止. 这里所谓按等概率分组，实际是尽可能接近，视具体信源而定.

例 1　试将下列信源编成二元码：

$$\mathscr{U} = [1,2,3,4,5,6],$$

其概率分布为

$$P = [0.4, 0.3, 0.1, 0.1, 0.06, 0.04]. \tag{2-16}$$

解　按哈夫曼(Huffman)编码法,取 $E_0 = [1]$, $E_1 = [2,3,4,5,6]$ 作为一线组;再将 E_1 分成等概率的两个小组 $E_{10} = [2]$, $E_{11} = [3,4,5,6]$;进一步,可分为 $E_{110} = [3,5]$, $E_{111} = [4,6]$;最后得信源 $\mathscr{U}$ 的二元码为 0,10,1100,1101,1110,1111,其平均码长为 $\bar{l} = 2.2\text{bit}$,这比信源的熵 $H = 2.1\text{bit}$ 稍大一点.

2. 申农编码法

这种方法是,首先,将信源概率按递减次序写出:$p_1 \geqslant p_2 \geqslant \cdots \geqslant p_q$. 其次,算出数列:$a_k = p_1 + p_2 + \cdots + p_{k-1}$; $k = 2, \cdots, q$, $a_1 = 0$. 第三,确定一列正整数 n_k, $k = 1,2,\cdots,q$,使

$$\log \frac{1}{p_k} \leqslant n_k < \log \frac{1}{p_k} + 1.$$

最后将 a_k 展成二进制小数(假定编制二元码),取前 n_k 位作为 p_k 对应的二元码字.

例 2　设信源概率分布如(2-16)式所示,试给出申农二元码.

解　$(p_1, p_2, p_3, p_4, p_5, p_6) = (0.4, 0.3, 0.1, 0.1, 0.06, 0.04)$,

$(a_1, a_2, a_3, a_4, a_5, a_6) = (0, 0.4, 0.7, 0.8, 0.9, 0.96)$,

$(n_1, n_2, n_3, n_4, n_5, n_6) = (2, 2, 4, 4, 5, 5)$,

将 a_k 展成二进小数,分别取前 n_k 位,即得申农二元码为

00,01,1011,1100,11100,11110,

其平均码长为　$\bar{l} = 2.7\text{bit}.$

3. 法诺编码法

这种方法是,首先将信源概率 $p(u)$, $u \in \mathscr{U}$ 由大到小排列;第二步是将末尾两个最小的概率相加,形成一个新的概率分布 $p_1(u)$, $u \in \mathscr{U}_1$,再按大小重新排序;最后将 p_1 的两个最小概率相加,形成 p_2 分布也按大小排序. 如此逐步简化信源,最后变成二元(两点)分布. 编码过程是先将最后的二元分布概率分别对应二元码符 0 及 1,倒回前一个三元分布,各编为 0,10 及 11(这里假定符号 1 对应的概率是由两个小概率合成的). 如此编号逐步退回原始分布,就编成了最后的二元法诺(Fano)码.

例 3　将(2-16)式的信源分布编成法诺码如表 2-1 所示.

表 2-1

$\mathscr{U}$	$p(u)$	$f(u)$	$p_1(u)$	$f_1(u)$	$p_2(u)$	$f_2(u)$	$p_3(u)$	$f_3(u)$	$p_4(u)$	$f_4(u)$
u_1	0.4	1	0.4	1	0.4	1	0.4	1	0.6	0
u_2	0.3	00	0.3	00	0.3	00	0.3	00	0.4	1
u_3	0.1	011	0.1	011	0.2	010	0.3	01		
u_4	0.1	0100	0.1	0100	0.1	011				
u_5	0.06	01010	0.1	0101						
u_6	0.04	01011								

定义 2　设对于给定信源 $[p(u), u \in \mathscr{U}]$,其熵为 $H(U)$,信源编码为 $f: \mathscr{U} \to \mathscr{B}^*$,则

$$\eta \xlongequal{\text{def}} \frac{H(U)}{\log|\mathscr{X}|} / El(f(U)) \tag{2-17}$$

称为码f的效率，其中$l(x)$为码字x的分量个数，即x的码长；称$1-\eta$为码f的冗长度.

综上所述，三种经典编码法各有千秋.法诺码的效率较申农码为高，哈夫曼码的效率最高；申农码是唯一定形的，其他两种码形均不唯一确定，即对于同一给定信源，按同一编码法(法诺或哈夫曼)，却可编制出不同形式的码来，并且是平均码长相等的唯一可译码.

紧致码是唯一可译，且是平均长最短的码.可以证明，哈夫曼码是紧致码.

2.3 带价值码

2.3.1 一般离散信源的熵率

现在考虑将平均码长定理3推广到一般有记忆离散信源的场合，并使用比平均码长更加广义的性能指标——“码的价值”.

一般有记忆信源，不具备公共熵，代替它的是熵率.

定义3 设$U=[U_i, i=1,2,\cdots]$是具有有限消息集合$\mathscr{U}$的任意离散信源(不限于无记忆)，$|\mathscr{U}|<\infty$.若下列极限存在：

$$\bar{H}(U) \xlongequal{\text{def}} \lim_{k\to\infty}\frac{1}{k}H(U^k) \qquad (U^k=(U_1,\cdots,U_k)), \tag{2-18}$$

则称之为信源U的熵率.

无记忆信源的熵率就是其公共分布的熵.

一个信源称为平稳的，如果$U_i, U_{i+1},\cdots,U_{i+k}$的联合分布与$i$无关，$k=0,1,2,\cdots$.

定理6 对于平稳信源，$[a_k=H(U^k)/k, k=1,2,\cdots]$是非增序列，因而熵率$\bar{H}(U)$总存在，且有

$$\bar{H}(U)=\lim_{k\to\infty}H(U_k/(U_1,\cdots,U_{k-1})). \tag{2-19}$$

2.3.2 码的价值

定义4 设码符集为$\mathscr{X}$，若每个信号元$x\in\mathscr{X}$，对应一正数$C(x)>0$，则称之为信号x的价值.

在背景上，价值可以理解为度量传递该信号或存储信号所需的时间或空间位置.信号所耗的时空单元，在经费上是有“价值”的.数学上，价值函数$C(x)$就是定义在码符集$\mathscr{X}$上的一个正实值函数.

定义5 设码符列为$x^n=(x_1,x_2,\cdots,x_n)$，则

$$C(x^n) \xlongequal{\text{def}} \sum_{i=1}^{n}C(x_i) \tag{2-20}$$

称为码符列x^n的价值.

特别地，当 $C(x)=1$ 时，$C(x^n)$ 即为码符列 x^n 的长度 n.

定义 6　若信源为$[U_i, i=1,2,\cdots]$，编码为 $f:\mathscr{U}^k\to\mathscr{X}^*$，则

$$\overline{C}(f)\xlongequal{\text{def}}\frac{1}{k}E[C(f(U^k))] \tag{2-21}$$

称为 k 长消息码的每个信源字母的平均价值.

特别地，当 $C(x)=1$ 时，$\overline{C}(f)=\bar{l}(f)$ 就是平均码长. 可见"平均价值"是"平均码长"概念的广义化.

设码符集为 $\mathscr{X}$，定义于其上的价值函数 $C(x)>0, x\in\mathscr{X}$，则方程

$$\sum_{x\in\mathscr{X}}\exp[-\alpha C(x)]=1 \tag{2-22}$$

存在唯一正根 α_0，且对于任意取值于 $\mathscr{X}$ 中的随机变量 X，均有

$$\frac{H(X)}{EC(X)}\leqslant\alpha_0, \tag{2-23}$$

其中等号成立，当且仅当 X 的分布为

$$p(x)=\exp\{-\alpha_0 C(x)\},\quad x\in\mathscr{X}. \tag{2-24}$$

2.3.3　平均价值定理

定理 7　对于任意离散信源$[U_i, i=1,2,\cdots]$，设(f,φ) 是 k 长消息的任一分组码

$$f:\mathscr{U}^k\to\mathscr{X}^*,\quad \varphi:\mathscr{X}^*\to\mathscr{U}^k,$$

且满足条件

$$\frac{1}{k}Ed[U^k,\varphi(f(U^k))]<\epsilon\leqslant\frac{1}{2}, \tag{2-25}$$

则有

$$\overline{C}(f)\geqslant\frac{H(U^k)}{k\alpha_0}-\frac{1}{\alpha_0}[\epsilon\log(|\mathscr{U}|-1)+H(\epsilon)]-\frac{[\log(dk)]^+}{k\alpha_0}, \tag{2-26}$$

其中 α_0 是方程(2-22) 的正根，

$$d=\frac{\mathrm{e}\cdot\log|\mathscr{U}|}{\alpha_0\cdot\min\limits_{x\in\mathscr{X}}C(x)}.$$

进一步，若 f 的值域唯一可译，则(2-26) 式中最后一项可去掉，这样对于任一唯一可译码 f(其值域唯一可译且 φ 等于 f 的逆)，有

$$\overline{C}(f)\geqslant\frac{H(U^k)}{k\alpha_0}. \tag{2-27}$$

此外，对于任意 k，存在一前束码，它满足

$$\overline{C}(f)<\frac{H(U^k)}{k\alpha_0}+\frac{1}{k}\max_{x\in\mathscr{X}}C(x). \tag{2-28}$$

需要指出，当 $C(x)=1$ 时，$\alpha_0=\log|\mathscr{X}|$，定理 7 是变长编码定理(定理 3) 的简单推广；若再假定信源是无记忆的，则二者归一.

推论1　在定理 7 的条件下，若进一步假定信源有熵率 $\overline{H}(U)$，则对于任意 $\delta>0$，存在 $\epsilon>0$ 及 k_0，使消息长为 $k\geqslant k_0$，且对于满足(2-25) 式的每个码 f，有

$$\bar{C}(f) > \frac{\bar{H}(U)}{\alpha_0} - \delta. \tag{2-29}$$

此外，对于任意 $\delta > 0$，及充分大的 k，存在前束码，满足

$$\bar{C}(f) < \frac{\bar{H}(U)}{\alpha_0} + \delta. \tag{2-30}$$

进一步，对于平稳信源及唯一可译码，下界可加强为

$$\bar{C}(f) \geqslant \frac{\bar{H}(U)}{\alpha_0}. \tag{2-31}$$

注：若将 $C(x)$ 理解为传递信号 x 的费用，则本推论 1 表明，在长消息传送中，平均每传递一个消息字母所需费用大致为 $\bar{H}(U)/\alpha_0$；若把熵率理解为每信源字母具有的信息量，则可将 α_0 解释为每单位费用最多能传递(或存储)的信息量.

2.3.4　信息统计问题

作为信源编码理论的一个应用，考虑这样一个"**信息统计**"问题：根据独立抽取的 k 个样本$(u_1, u_2, \cdots, u_k)$，试判断它原来的分布(母体)是 $P = \{p(u), u \in \mathscr{U}\}$，还是 $Q = [q(u), u \in \mathscr{U}]$，$|\mathscr{U}| < \infty$.

这是个普通的假设检验问题.解决的办法是，选定一个适当的集合 $A \subset \mathscr{U}^k$，并作判定：若样本$(u_1, \cdots, u_k) \in A$，则接受 P；否则，接受 Q.详细地说，要选择 A，使得一方面当 P 为真时，"弃真"的概率小于预定限度 ϵ，即 $P(\bar{A}) \leqslant \epsilon$ 或 $P(A) \geqslant 1 - \epsilon$；另一方面，当 Q 为真时，使得"存伪"的概率 $Q(A)$ 达到最小，即

$$\beta(k, \epsilon) \xlongequal{\text{def}} Q(A) = \min[Q(A'): A' \subset \mathscr{U}^k, P(A') \geqslant 1 - \epsilon].$$

需要注意，这里 A 的选择原则，对于 P 与 Q 是不对称的.

一般统计方法是，选取 A 使在保证"弃真"概率不超过 ϵ 条件下，使"存伪"概率尽可能小，但对于这个所谓存伪概率究竟能小到何种地步，就不得而知了.

运用编码定理的思想可以证明：对于任意 $0 < \epsilon < 1$，有

$$\lim_{k \to \infty} \frac{1}{k} \log \beta(k, \epsilon) = - D(P \,/\!/\, Q), \tag{2-32}$$

其中 $D(P /\!/ Q)$ 称为分布 P 相对 Q 的"**信息散度**"，用它度量 Q 不同于 P 的程度.

按定义，

$$D(P \,/\!/\, Q) \xlongequal{\text{def}} \sum_{u \in \mathscr{U}} p(u) \log \frac{p(u)}{q(u)}, \tag{2-33}$$

由基本不等式有 $D(P /\!/ Q) \geqslant 0$.它越大，一次观察所能获取的分辨 P 与 Q 的"信息"就越多.

2.4　具保真度码

2.4.1　失真度

(1) 实际获取的原始信息，如卫星图像等，一方面数据量很大，另一方面还混

杂着干扰信号,于是需经适当处理才便于传递和存储.常用的方法是,利用滤波技术去除污染,同时还要将数据进行压缩变换以减少冗余信息.

原始数据经过编码压缩掉一定的消息后,再通过译码来复现消息,这一编一译,便产生了信息"失真"的问题.

具保真度的信源编码,就是一方面要最大限度地压缩信源消息数据量,另一方面还要使失真不能太严重.也就是说,在一定保真度下来尽可能地压缩消息量.那么,什么是"**失真度**"呢?

设 $\mathscr{U}$ 为信源的消息集合,$|\mathscr{U}|<\infty$,而 $\mathscr{V}$ 为其复制信号的集合(或称目标集合),$|\mathscr{V}|<\infty$.一般说来,所谓**失真度**,就是定义在 $\mathscr{U}\times\mathscr{V}$ 上的一个非负二元实值函数 $d(u,v)\geqslant 0$.它表示用信号 $v\in\mathscr{V}$ 来复制消息 $u\in\mathscr{U}$ 所产生的失真度量.常选择 d 使其较小值代表小的失真,且假定,对于每个 $u\in\mathscr{U}$,至少有一个 $v\in\mathscr{V}$ 使 $d(u,v)=0$;不然,可令

$$d'(u,v)=d(u,v)-\min_{v\in\mathscr{V}}d(u,v),$$

这样 d' 满足上述要求.

特别地,当 $\mathscr{U}=\mathscr{V}$ 时,可取失真度为

$$d(u,v)=\begin{cases}0, & \text{当 } u=v;\\ 1, & \text{当 } u\neq v.\end{cases}$$

当 $\mathscr{U}=\mathscr{V}\subset\mathbf{R}^1$ 为实数时,还可取

$$d(u,v)=(u-v)^2.$$

这两例中,前者要求严格地复现消息,后者则表示随信号误差幅度的增大而使失真严重化.

(2) 设随机变量 U 具有概率分布 $[p(u),u\in\mathscr{U}]$.如果对于"随机消息"U,可选择取值于信号集合 $\mathscr{V}$ 中的"随机信号"V 作为其"复制品",而 V 的概率分布 $p(v)$,$v\in\mathscr{V}$ 可以规定为

$$p(v)\xlongequal{\text{def}}\sum_{u\in\mathscr{U}}p(u)Q(v/u),\tag{2-34}$$

这里对不同随机变量的概率函数的标示符号不加区分,都用 p,只是用 $p(u)$ 或 $p(v)$ 来区分是 U 或 V 的分布,其中 $Q(v/u)$ 是 $\mathscr{U}$ 到 $\mathscr{V}$ 上的转移概率函数(条件概率),

$$\sum_{v\in\mathscr{V}}Q(v/u)=1,\quad\forall\, u\in\mathscr{U},\quad Q(v/u)\geqslant 0,\tag{2-35}$$

则在 $[\mathscr{U},p(u),\mathscr{V}]$ 给定的条件下,对 V 的选择就等同于对 $[Q(v/u),(u,v)\in\mathscr{U}\times\mathscr{V}]$ 的选择.一旦这个转移概率确定,便得 U 与 V 的联合分布为

$$p(u,v)\xlongequal{\text{def}}p(u)Q(v/u).\tag{2-36}$$

于是可定义**平均失真度**为

$$d(Q)\xlongequal{\text{def}}\sum_{u,v}p(u)Q(v/u)d(u,v)=Ed(U,V).\tag{2-37}$$

另外,还可得到 U 与 V 间的互信息为

$$I(U;V) \equiv I(P;Q) \xlongequal{\text{def}} \sum_{u,v} p(u)Q(v/u)\log\frac{Q(v/u)}{p(v)}. \tag{2-38}$$

2.4.2 率失真函数

(1) 直觉上,信源消息 U 被复制成信号 V 之后,其"失真度"越小,它们间的"互信息"就会越大;反之,"信息压缩"越大,"复制品"失真必然严重.这就是说,所谓保真压缩信源编码问题,实际上是在复制信号 V 与原始消息 U 之间的互信息 $I(U;V)$ 与它们之间的失真度 $Ed(U,V)$ 这二者中进行协调选择的问题.在一定失真度限制下,使互信息最小,就产生了最大压缩,这种压缩的性能指标就称之为"率失真函数".其严格定义如下.

定义7 设信源为$[U_j, j=1,2,\cdots]$,每个分量 U_j 皆取值于消息集 $\mathscr{U}$;对于任意 k,将 k 长消息 $U^k=(U_1,U_2,\cdots,U_k)$ 复制为 k 长随机信号 $V^k=(V_1,\cdots,V_k)$,每个 V_j 皆取值于信号集 $\mathscr{V}$ 中.在 $\mathscr{U}\times\mathscr{V}$ 上给定失真度 $d(u,v)\geqslant 0$,则对于固定 k,

$$R_k(\delta) \xlongequal{\text{def}} \min_{V^k:Ed(U^k,V^k)\leqslant k\delta} I(U^k;V^k) \tag{2-39}$$

称为 k 率失真函数,其中最小值是对所有这样的随机向量 V^k 而取的,其使 $Ed(U^k,V^k)\leqslant k\delta$.按(2-37)式及(2-38)式的记号,也可将(2-39)式写为

$$R_k(\delta) \xlongequal{\text{def}} \min_{Q:d(Q)\leqslant k\delta} I(P;Q), \tag{2-40}$$

这里 P 是 U^k 的概率分布,而 Q 则是已知 U^k 下,V^k 的转移(条件)概率.满足限制 $d(Q)\leqslant k\delta$ 的 Q 称之为许用转移概率.

定义8 设信源列为$[U_j, j=1,2,\cdots]$,则

$$R(\delta) \xlongequal{\text{def}} \inf_{k\geqslant 1}\frac{1}{k}R_k(\delta) \tag{2-41}$$

称为信源列的率失真函数.

关于率失真函数的几点注释:

1) 定义(2-40)式的极小值一定存在,因为 $I(P;Q)$ 可视为诸元 $Q(v/u)$ 的多元连续函数,其元素个数为 $m=|\mathscr{U}|^k\cdot|\mathscr{V}|^k$;而限制条件 $d(Q)\leqslant k\delta$ 实质是 m 维闭区域.

2) $R_k(\delta)$ 的定义域为 $\delta\geqslant\delta_{\min}$,这里

$$\delta_{\min} \xlongequal{\text{def}} \sum_{u\in\mathscr{U}} P(u)\cdot\min_{v\in\mathscr{V}} d(u,v). \tag{2-42}$$

3) 率失真函数 $R(\delta)$ 的信息含义是,允许平均失真 δ 时,复现每一消息字母所需的最小比特数(假定用二元码,对数以2为底).

(2) 率失真函数具有下列性质:

1° $R_k(\delta)$ 是 $\delta\geqslant\delta_{\min}$ 的单减下凸函数;

2° $R_k(\delta)$ 是 $\delta\geqslant\delta_{\min}$ 的连续函数;

3° $R_k(\delta)=0$,当且仅当 $\delta\geqslant\delta_{\max}$,这里

$$\delta_{\max} \xlongequal{\text{def}} \min_{v\in\mathscr{V}}\sum_{u\in\mathscr{U}} p(u)d(u,v); \tag{2-43}$$

4° 当$\delta_{\min} \leqslant \delta \leqslant \delta_{\max}$时，$R_k(\delta)$严格单减，从而有

$$R_k(\delta) = \min_{d(Q)=k\delta} I(P;Q), \quad \delta_{\min} \leqslant \delta \leqslant \delta_{\max}; \tag{2-44}$$

5° 对于离散无记忆信源，有

$$R_k(\delta) = kR_1(\delta),$$
$$R(\delta) = R_1(\delta), \quad \delta \geqslant \delta_{\min}. \tag{2-45}$$

2.4.3 保真信源编码定理

设离散无记忆信源$[U_j, j = 1,2,\cdots]$，具有公共分布$p(u), u \in \mathscr{U}, |\mathscr{U}| < \infty$；复制信号集合$\mathscr{V}, |\mathscr{V}| < \infty$；在$\mathscr{U} \times \mathscr{V}$上给定失真度为$d(u,v) \geqslant 0$. 另外，取$\mathscr{V}^k$的一个子集$C$，

$$C = [v_i \in \mathscr{V}^k; \quad i = 1,2,\cdots,M],$$

并称之为k长码字集合，简称码C，其中码字个数为$M = |C|$. 对于每一k长消息$u \in \mathscr{U}^k$，记号

$$d(u,C) \xlongequal{\text{def}} \min_{v \in C} d(u,v), \tag{2-46}$$

表示u到码C的最小失真度. 对于任何$u \in \mathscr{U}^k$，C中总有这样一个码字，记之为$v(u)$，满足

$$d(u,v(u)) = d(u,C). \tag{2-47}$$

由此定义的$v(u)$可视为$\mathscr{U}^k$到$\mathscr{V}^k$的一个翻码. 对于k长的随机消息$U^k = (U_1, U_2, \cdots, U_k)$，定义码$C$的平均失真度为

$$d(C) \xlongequal{\text{def}} \frac{1}{k} Ed(U^k, C) = \frac{1}{k} \sum_{u \in \mathscr{U}^k} p(u) d(u, v(u)). \tag{2-48}$$

码C的信息压缩率为

$$R_t \xlongequal{\text{def}} \frac{1}{k} \log M. \tag{2-49}$$

现在的问题是，在平均失真$d(C)$不超过允许限度$\delta(d(C) \leqslant \delta)$时，信源压缩率$R_t$的限度是什么？

由于对于k长随机消息$U^k = (U_1, U_2, \cdots, U_k)$及任一随机码字$V^k = (V_1, V_2, \cdots, V_k)$，总有

$$I(U^k;V^k) \leqslant H(V^k) \leqslant \log M, \tag{2-50}$$

假定随机向量U^k与V^k间的转移概率分布Q使平均失真度不超过δ，即

$$d(Q) = Ed(U^k, V^k) \leqslant k\delta,$$

则有

$$R_k(\delta) \leqslant I(U^k;V^k) \leqslant \log M,$$

从而有

$$R(\delta) \leqslant \frac{1}{k} R_k(\delta) \leqslant \frac{1}{k} \log M.$$

由此便得

$$R_s \geqslant R(\delta), \quad M \geqslant 2^{kR(\delta)}. \tag{2-51}$$

这表明,在平均失真度不超过 δ 的情况下,k 长消息信源至少要用 $M \geqslant 2^{kR(\delta)}$ 个码字来复制;或者说,当码字个数少于 $2^{kR(\delta)}$ 时,平均失真要超过限度 δ.

试问有无这样的码 C,使其中码字个数 $M \geqslant 2^{kR(\delta)}$,且平均失真度又不超过给定的 δ,即 $d(C) \geqslant \delta$?答案是,这两条要同时满足不易,但近似成立的码 C 确实存在,这就是下面的申农保真信源编码定理的结论.

定理 8 设具失真度的无记忆信源

$$\mathscr{S} = \{[\mathscr{U}, P(u)], \mathscr{V}, d(u,v)\}, \tag{2-52}$$

若给定 $\delta \geqslant \delta_{\min}$,则对于任意 $\epsilon > 0, \rho > 0$,当 k 充分大时,一定存在 k 长信源码 C,它具有 M 个码字,且满足下列两个条件:

1° $M < 2^{k\cdot(R(\delta)+\epsilon)}$;

2° $d(C) < \delta + \rho$.

注:上述结果对码字数目的指标 $R(\delta)$ 多了个 ϵ,也就是此码的压缩率 R_s 为

$$R(\delta) \leqslant R_s < R(\delta) + \epsilon; \tag{2-53}$$

另外,失真度也未严格限于 δ,而多了个 $\rho > 0$.

精细分析可知,这两个界多出的"残量"ϵ 和 ρ,在有限形式(不取极限)下,同时去除是很难的;任意去掉一个是可能的,由此引出推论 2.

推论2 在定理8 的条件下,有

1° 对于任意 $\delta > \delta_{\min}, \epsilon > 0$,当 k 充分大时,存在 k 长信源码 C,同时满足

$$R_s < R(\delta) + \epsilon, \quad d(C) < \delta;$$

2° 对于任意 $\delta_{\min} \leqslant \delta \leqslant \delta_{\max}$ 及 $\rho > 0$,当 k 充分大时,存在 k 长信源码 C,使

$$R_s < R(\delta), \quad d(C) < \delta + \rho.$$

需要指出,推论 2 中两种结果,本质上是在严格(不等式)条件 $\delta > \delta_{\min}$ 及 $\delta_{\min} \leqslant \delta < \delta_{\max}$ 下,调整两类指标"压缩限"与"失真度",以放宽一个为代价,而换取另一个的精确限度的结果.

2.4.4 率失真函数的计算

从上述保真信源编码定理可知,率失真函数 $R(\delta)$ 是在信源码平均失真不超过 δ 下,信息压缩率的下限标志.因而给定信源及失真度时,如何计算其率失真函数 $R(\delta)$,就是个关键问题.

例 4 设具失真度的无记忆信源为

$$\mathscr{S} = \{[\mathscr{U}, p(u)], \mathscr{V}, d(u,v)\},$$

其中 $\mathscr{U} = [0,1] = \mathscr{V}, p(0) = p \leqslant \frac{1}{2}, p(1) = 1 - p = q,$

$$d(u,v) = \begin{cases} 0, & \text{当 } u = v; \\ 1, & \text{当 } u \neq v. \end{cases}$$

试求该信源的率失真函数 $R(\delta)$.

解 先确定 $R(\delta)$ 的定义域及非零值范围,这由 $\delta_{\min}$ 及 $\delta_{\max}$ 给出.据(2-42)式及(2-43)式,可得

$$\delta_{\min} = 0, \quad \delta_{\max} = \min\{p, q\} = p,$$

再由(2-44)式,有

$$R(\delta) = \min_{d(Q)=\delta} I(P;Q) = \min_{Ed(U,V)=\delta} I(U;V) \quad (0 \leqslant \delta < p).$$

由互信息的性质及法诺不等式,有

$$\begin{aligned} I(U;V) &= H(U) - H(U/V) = H(p) - H(U/V) \\ &\geqslant H(p) - H(p_e). \end{aligned}$$

其中 $H(p) = -p\log p - q\log q$,为熵函数($q = 1-p$);$p_e \stackrel{\text{def}}{=} P[U \neq V]$ 为误差概率.

注意到 $d(Q) = Ed(U,V) = p_e$,便有

$$R(\delta) \geqslant H(p) - H(\delta), \quad 0 \leqslant \delta < p \leqslant \frac{1}{2}.$$

另外,可取转移概率 Q,使 $H(U/V) = H(\delta)$,为此,只要满足下列条件

$$p(u/v) = \begin{cases} \delta, & \text{当 } u \neq v; \\ 1-\delta, & \text{当 } u = v, \end{cases}$$

这样的 Q 可取为

$$\begin{aligned} Q(0/0) &= \lambda(1-\delta)/p, \quad Q(1/0) = (1-\lambda)\delta/p, \\ Q(0/1) &= \lambda\delta/q, \quad Q(1/1) = (1-\lambda)(1-\delta)q, \\ \lambda &= (p-\delta)/(1-2\delta). \end{aligned}$$

它恰好是一允许转移概率,同时满足

$$I(U;V) = H(p) - H(\delta), d(Q) = \delta.$$

于是又得

$$R(\delta) \leqslant I(U;V) = H(p) - H(\delta).$$

最后

$$R(\delta) = \begin{cases} H(p) - H(\delta) & (\text{当 } 0 \leqslant \delta \leqslant p \leqslant \frac{1}{2}); \\ 0 & (\text{当 } \delta \geqslant p). \end{cases}$$

这就是两点信源 $\mathscr{S}_1$的率失真函数.

此例显示,率失真函数的计算问题,即使对于简明的信源也是不容易的事情.一般有下列定理.

定理 9 若给定无记忆信源$\{[\mathscr{U}, p(u)], \mathscr{V}, d(u,v)\}$,则其率失真函数为

$$R(\delta) = \max_{s \leqslant 0, \lambda \in \Lambda_s} \{s\delta + \sum_{u \in \mathscr{U}} p(u) \log \lambda(u)\}, \tag{2-54}$$

其中

$$\Lambda_s \stackrel{\text{def}}{=} \{\lambda(u), u \in \mathscr{U}: \sum_{u \in \mathscr{U}} \lambda(u) p(u) \exp[sd(u,v)] \leqslant 1, \forall v \in \mathscr{V}\}. \tag{2-55}$$

由定理 9 可得率失真函数的计算步骤如下:

(1) 求协变函数 $\lambda(u), u \in \mathscr{U}$.

$$\sum_{u \in \mathscr{U}} \lambda(u) p(u) \exp[sd(u,v)] = 1 \quad (\forall v \in \mathscr{V}). \tag{2-56}$$

需要注意,$\mathscr{U}$与$\mathscr{V}$皆为有限元素的集合,故方程(2-56)式实际上是具有$|\mathscr{U}|$个未知数$\lambda(u)$,$u\in\mathscr{U}$,及$|\mathscr{V}|$个方程的方程组.

(2) 确定$\mathscr{V}$上的概率分布$p(v)$.

$$\sum_{v\in\mathscr{V}}p(v)\exp[sd(u,v)]=\frac{1}{\lambda(u)}\quad(\forall u\in\mathscr{U}),\tag{2-57}$$

此时,$\lambda(u)$与$p(v)$都是s的函数.

(3) 将求得的$\lambda(u)$与$p(v)$代入下式,便得转移概率$Q(v/u)$,它也是s的函数.

$$Q(v/u)=\lambda(u)p(v)\exp[sd(u,v)]\quad(u\in\mathscr{U},v\in\mathscr{V}).\tag{2-58}$$

(4) 由下式确定s作为δ的函数.

$$\sum_{u,v}p(u)Q(v/u)d(u,v)=\delta.\tag{2-59}$$

(5) 将$s=s(\delta)$及$\lambda(u)$作为δ的函数代入下式,便得率失真函数为

$$R(\delta)=s(\delta)+\sum_{u\in\mathscr{U}}p(u)\log\lambda(u),\tag{2-60}$$

其中参量s实际上是$R(\delta)$的斜率,即

$$s=\frac{\mathrm{d}R(\delta)}{\mathrm{d}\delta},\tag{2-61}$$

但在$R(\delta)$未知之前不能由此式求得s.

按上述算法,计算例4的$R(\delta)$,可得同样结果.进一步的例子可见参考文献3.

3 信道编码

3.1 噪声信道编码问题

3.1.1 信道与编码

设$\mathscr{X}$与$\mathscr{Y}$都是有限集合,$|\mathscr{X}|<\infty$,$|\mathscr{Y}|<\infty$.以$\mathscr{X}$为入口信号集,$\mathscr{Y}$为出口信号集的一个信道,是由一族条件概率$\{Q(y/x);x\in\mathscr{X},y\in\mathscr{Y}\}$来决定的,其中$Q(y/x)\geqslant 0$表示发送入口信号$x$时收到出口信号$y$的条件概率,简称$x$到$y$的转移概率,有

$$\begin{cases}\sum_{y\in\mathscr{Y}}Q(y/x)=1 & (\forall x\in\mathscr{X}),\\ Q(y/x)\geqslant 0 & (\forall(x,y)\in\mathscr{X}\times\mathscr{Y}).\end{cases}\tag{3-1}$$

该信道记为$[X,Q(y/x),\mathscr{Y}]$,也称Q为转移阵,它实质上是具有$|\mathscr{X}|$行及$|\mathscr{Y}|$列的随机矩阵,该矩阵每个元素取非负值,且每行之和为1.

具有入口集$\mathscr{X}$及出口集$\mathscr{Y}$的一个**信道码**,就是一对映射(f,g),其中f是把某有限集合$\mathscr{U}$映射到$\mathscr{X}$,称之为**翻码器**(简称**翻码**),

$$f: \mathscr{U} \to \mathscr{X};$$

而 g 是把 $\mathscr{Y}$ 映射到某一集合 $\mathscr{V}$,称之为译码器(简称译码),

$$g: \mathscr{Y} \to \mathscr{V};$$

$\mathscr{U}$ 称为消息集.信道码(f,g) 的任务是先将要发送的每一消息 $u \in \mathscr{U}$,翻成信道入口的一个码字 $x = f(u)$,经信道传递,在信道出口处输出信号 y,由于信道可能有噪声干扰(这体现在转移概率 $Q(y/x)$ 中),需对 y 经译码手续判断出所发送的信号及消息,而后送到用户(信宿)去,通常假定 $\mathscr{U} = \mathscr{V}$.

3.1.2 通信系统及误差概率

信源消息是随机发送的,若其概率分布给定为 $p(u), u \in \mathscr{U}$,则信源$[\mathscr{U}, p(u)]$、信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$ 及信道编码(f,g) 便组成一个通信系统,如图 3-1 所示,其中信源消息 u 及对应的码字 $x = f(u)$,相应的接收信号 y 和信宿消息 $v = g(y)$,这四个量都是随机样本,它们对应的随机变量及传递关系可表示为

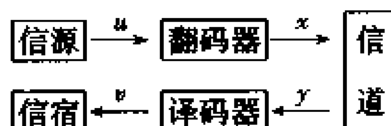


图 3-1

$$U \xrightarrow{f} X \xrightarrow{Q} Y \xrightarrow{g} V,$$

其联合概率分布为

$$p(u,x,y,v) = p(u)p(x/u)Q(y/x)p(v/y), \tag{3-2}$$

这里将翻码 f 及译码 g 的作用,也表成了转移概率 $p(x/u)$ 及 $p(v/y)$,并分别规定为

$$p(x/u) \xlongequal{\text{def}} \begin{cases} 1 & (\text{当 } x = f(u)), \\ 0 & (\text{当 } x \neq f(u)); \end{cases}$$

$$p(v/y) \xlongequal{\text{def}} \begin{cases} 1 & (\text{当 } v = g(y)), \\ 0 & (\text{当 } v \neq g(y)). \end{cases}$$

由此确定(U,V) 的联合分布 $p(u,v)$ 及通信误差概率 p_e 分别为

$$p(u,v) = \sum_{x,y} p(u,x,y,v) = p(u)Q(g^{-1}(v)/f(u)),$$

$$\forall (u,v) \in \mathscr{U} \times \mathscr{V}; \tag{3-3}$$

$$p_e = p_e(f,g) \xlongequal{\text{def}} P[U \neq V] = \sum_u p(u) \cdot p_{e,u}, \tag{3-4}$$

其中

$$p_{e,u} \xlongequal{\text{def}} \sum_{v \neq u} Q(g^{-1}(v)/f(u)) \quad (\forall u \in \mathscr{U}),$$

$$Q(g^{-1}(v)/f(u)) \xlongequal{\text{def}} \sum_{y: g(y)=v} Q(y/f(u)).$$

这里 $p_{e,u}$ 表示发送单个消息 u 时产生的误差概率;而 p_e 则是发送所有消息 $u \in \mathscr{U}$

之平均误差概率.特别地,对于等概率信源分布 $p(u)=1/|\mathscr{U}|$,有

$$p_e=\frac{1}{|\mathscr{U}|}\cdot\sum_{u\in\mathscr{U}}p_{e,u},\tag{3-5}$$

其中 $p_{e,u}$ 只与信道及编码(f,g)有关,与信源分布无关,它实际上是传递码字 $x=f(u)$ 时的条件误差概率,其最大值记为

$$p_e(m)=\max_{u\in\mathscr{U}}p_{e,u}\overset{\text{def}}{=\!=}\max_{u\in\mathscr{U}}\sum_{v\neq u}Q(g^{-1}(v)/f(u)).\tag{3-6}$$

所谓信道编码问题,就是选择编码(f,g),使可用消息的发送数目尽可能大(即码字个数尽可能多),同时还要求最大误差概率 $p_e(m)$ 尽量小.

3.1.3 无记忆信道

实际通信总是在一定时间内接连发送信号,对信道"多次利用".

定义1 设信道为$[\mathscr{X}_n,Q_n(y^n/x^n),\mathscr{Y}_n]$, $n=1,2,\cdots$.

若 $\mathscr{X}_n=\mathscr{X}^n$、$\mathscr{Y}_n=\mathscr{Y}^n$,且

$$Q_n(y^n/x^n)=\prod_{i=1}^{n}Q(y_i/x_i),$$

则称该信道为**无记忆信道**,其中

$$x^n=(x_1,x_2,\cdots,x_n)\in\mathscr{X}^n;\quad y^n=(y_1,y_2,\cdots,y_n)=\mathscr{Y}^n;$$

简记为$[\mathscr{X},Q(y/x),\mathscr{Y}]$.

无记忆信道所描述的实际背景是,当接连发送一列信号$(x_1,x_2,\cdots,x_n)$,并相继收到对应信号列$(y_1,y_2,\cdots,y_n)$时,y_1只依赖于x_1;y_2只依赖x_2而与x_1和y_1都无关;…;y_n只与x_n有关而与它前面发送的信号都无关.

理论上简单而非平凡的无记忆信道,就是**二元对称信道**$[\mathscr{X},Q(y/x),\mathscr{Y}]$,其中

$$\mathscr{X}=\mathscr{Y}=\{0,1\},$$

$$Q(y/x)=\begin{cases}\varepsilon & (\text{当 } x\neq y);\\ 1-\varepsilon & (\text{当 } x=y),\end{cases}\tag{3-7}$$

其中 $\varepsilon(0<\varepsilon<\frac{1}{2})$ 作为二元信道参数,表示误码率.

下面主要介绍 k 到 n 定长分组码(f,g)的编码定理,这里约定:$\mathscr{U}=\mathscr{V}$,且$\mathscr{X}=\mathscr{Y}$;$f:\mathscr{U}^k\to\mathscr{X}^n$, $g:\mathscr{Y}^n\to\mathscr{V}^k$;简写通信系统为

$$u^k\xrightarrow{f}x^n\xrightarrow{Q}y^n\xrightarrow{g}v^k.$$

3.2 信道容量与逆编码定理

3.2.1 信道容量

如果说具保真度的信源编码定理是在一定的"平均失真度"下,解决信源码字的"最小压缩数目"问题,那么信道编码定理则是在一定的"误差概率限度"下,解

决信道码字的"最大许用数目"问题.也就是说,对信源码而言,在一定保真度下,信源码字压缩数目越少越好;而信道码则要求在允许误差概率下,信道码字越多越好.前者的下限标志,是信源的率失真函数;而后者的上限标志,就是**信道容量**.两者异曲同工.

给定信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$,并在入口信号集$\mathscr{X}$上指定任一概率分布$p(x)$, $x \in \mathscr{X}$,便产生二随机变量X与Y,它们的联合分布为

$$p(x,y) \xlongequal{\text{def}} p(x)Q(y/x) \qquad (\forall (x,y) \in \mathscr{X} \times \mathscr{Y}).$$

由此得互信息为

$$I(P;Q) = I(X;Y) = \sum_{x,y} p(x)Q(y/x)\log \frac{Q(y/x)}{\sum\limits_{x' \in \mathscr{X}} p(x')Q(y/x')}, \qquad (3\text{-}8)$$

其中Q是信道标志,作为预定量;而互相信$I(P;Q)$作为入口分布P的函数,是上凸函数(见(1-21)式),且是P的变动区域D上的连续函数,这里$P = [p(x), x \in \mathscr{X}]$,记

$$D = [P: p(x) \geqslant 0, \sum_{x} p(x) = 1].$$

因而$I(P;Q)$在D上有极大值,

$$C \xlongequal{\text{def}} \max_P I(P;Q). \qquad (3\text{-}9)$$

定义 2 互信息$I(P;Q)$关于入口分布P的最大值C称为给定信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$的**信道容量**.

它是信道的特征数,只作为信道转移阵Q的函数,而与入口分布P无关.

像率失真函数一样,信道容量C的计算问题,本质上仍是多元函数的条件极值问题,一般并不容易计算.下面举例说明.

例 1 试求**二元对称信道**$[\mathscr{X}, Q(y/x), \mathscr{Y}]$的容量$C$,假定信道参数为$\epsilon$, $0 < \epsilon < \frac{1}{2}$.

解法 1 任意给定入口分布$p(x)$, $x \in [0,1]$,则有

$$I(X;Y) = H(Y) - H(Y/X) = H(Y) - H(\epsilon),$$

这里$H(\epsilon) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$为熵函数.于是

$$C = \max_P H(Y) - H(\epsilon).$$

另外,由熵的性质,得

$$\max_P H(Y) = \log 2,$$

为此,要求出口分布为等概率分布.又因信道的对称性,只要取入口分布也为等概率即可.于是得结果

$$C = \log 2 - H(\epsilon) = \log 2 + \epsilon\log\epsilon + (1-\epsilon)\log(1-\epsilon)$$

$$(0 < \epsilon < \frac{1}{2}). \qquad (3\text{-}10)$$

解法 2 记任一入口分布为

$$\alpha = P\{X=0\}, \bar{\alpha} = 1-\alpha = P\{X=1\},$$

这样,出口分布可写为

$$\beta = P[Y=0] = \sum_{x=0}^{1} p(x)Q(0/x) = \alpha\bar{\epsilon} + \bar{\alpha}\epsilon,$$

$$\bar{\beta} = 1-\beta = P[Y=1] = \sum_{x=0}^{1} p(x)Q(1/x) = \alpha\epsilon + \bar{\alpha}\bar{\epsilon}.$$

入口分布与出口分布二者关系可用矩阵表为

$$[\beta, \bar{\beta}] = [\alpha, \bar{\alpha}] \begin{bmatrix} \bar{\epsilon} & \epsilon \\ \epsilon & \bar{\epsilon} \end{bmatrix},$$

于是互信息可写为

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y/X) = H(\beta) - H(\epsilon) \\ &= H(\alpha\bar{\epsilon} + \bar{\alpha}\epsilon) - H(\epsilon). \end{aligned}$$

由此可见,当 $\alpha = \bar{\alpha} = \frac{1}{2}$ 时,上式有极大值

$$C = \max_{0 \le \alpha \le 1} I(X;Y) = \log 2 - H(\epsilon).$$

上述两种算法,结果都一样.

3.2.2 逆编码定理

给定无记忆信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$.对任一信源分布 $p(u)$, $u = (u_1, u_2, \cdots, u_k) \in \mathscr{U}^k$,考虑 k 到 n 定长分组码(f, g),

$$f: \mathscr{U}^k \to \mathscr{X}^n, \quad g: \mathscr{Y}^n \to \mathscr{V}^k;$$

由此在 $\mathscr{U}^k, \mathscr{X}^n, \mathscr{Y}^n$ 及 $\mathscr{V}^k = \mathscr{U}^k$ 上形成四个随机向量,按通信顺序构成马氏链:

$$U^k \xrightarrow{f} X^n \xrightarrow{Q} Y^n \xrightarrow{g} V^k. \tag{3-11}$$

规定此通信系统的传信率为

$$R_H \xlongequal{\text{def}} \frac{H(U^k)}{n}, \tag{3-12}$$

其中 $H(U^k)$为产生 k 长消息之信源的申农熵; R_H 表示每个信道符号平均传递的信息量.

上述通信系统的误差概率为

$$\begin{aligned} p_e &\xlongequal{\text{def}} P\{U^k \ne V^k\} = P\{\bigcup_{i=1}^{k}(U_i \ne V_i)\} \\ &\geqslant \max_{1 \le i \le k} P\{U_i \ne V_i\} \geqslant \frac{1}{k}\sum_{i=1}^{k} p_{e,i} \xlongequal{\text{def}} \langle p_e \rangle, \end{aligned}$$

其中

$$p_{e,i} = P\{U_i \ne V_i\} \quad (i = 1, 2, \cdots, k). \tag{3-13}$$

定理1 (逆编码定理) 设无记忆信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$ 的容量为 C,则对于任一 k 到 n 定长分组码(f, g),

$$f: \mathscr{U}^k \to \mathscr{X}^n, g: \mathscr{Y}^n \to \mathscr{V}^k;$$

及任一信源分布 $p(u), u \in \mathscr{U}^k$,有

$$\langle p_e\rangle\log(|\mathscr{U}|-1)+H(\langle p_e\rangle)\geqslant\frac{H(U^k)}{k}-\frac{n}{k}C$$

$$=\frac{n}{k}(R_H-C). \tag{3-14}$$

此外,当 $p_e\leqslant\frac{1}{2}$ 时,有

$$p_e\log(|\mathscr{U}|-1)+H(p_e)\geqslant\frac{n}{k}(R_H-C). \tag{3-15}$$

其中 $H(p_e)=-p_e\log p_e-(1-p_e)\log(1-p_e)$, $R_H=H(U^k)/n$, $|\mathscr{U}|$ 为 $\mathscr{U}$中元素的个数;对数以任何 $a>1$ 为底.

定理 1 表明,当保持传信率 R_H 大于信道容量 C 时,无论怎样选择信道编码(f, g),都不能使误差概率任意小!

3.3 具价值的信道编码

3.3.1 价值容量函数

定义 3 若对无记忆信道$[\mathscr{X}, Q(y/x), \mathscr{Y}]$ 的每个入口信号 $x\in\mathscr{X}$,指定一非负数 $b(x)\geqslant 0$,则称之为 x 的价值.

定义 4 设 n 长入口信号向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)\in\mathscr{X}^n$,则

$$b(\boldsymbol{x})\stackrel{\text{def}}{=\!=}\sum_{i=1}^{n}b(x_i)\quad(i=1,2,\cdots,n), \tag{3-16}$$

称为 n 长入口信号向量 $\boldsymbol{x}$ 的向量价值.

定义 5 设 n 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$,则

$$E[b(\boldsymbol{X})]\stackrel{\text{def}}{=\!=}\sum_{\boldsymbol{x}\in\mathscr{X}^n}b(\boldsymbol{x})p(\boldsymbol{x})=\sum_{i=1}^{n}E[b(X_i)], \tag{3-17}$$

称为随机向量 X 的平均价值.

定义 6 设信道转移阵为 Q,则

$$C_n(\beta)\stackrel{\text{def}}{=\!=}\max_{\boldsymbol{X}:E[b(\boldsymbol{X})]\leqslant n\beta}I(\boldsymbol{X};\boldsymbol{Y})\stackrel{\text{def}}{=\!=}\max_{P:b(P)\leqslant n\beta}I(P;Q), \tag{3-18}$$

称为信道的 n 阶价值容量函数,其中 P 是 n 阶入口向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ 的任一分布概率函数;$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)$ 是相应的出口向量.(3-18) 式中最大值是对所有入口向量 $\boldsymbol{X}$,即变动入口分布 P 而取的,其使 $b(P)=E[b(\boldsymbol{X})]\leqslant n\beta$,即 n 阶入口分布的平均价值不超过 $n\beta$;称这样的分布 P 为β允许试验入口分布(简称β允许分布).

注:(1) $C_n(\beta)$ 的定义(3-18) 式中最大值一定存在,此因 $I(\boldsymbol{X};\boldsymbol{Y})=I(P;Q)$ 在 Q 一定下,是入口分布 P 的连续上凸函数,极值条件不等式 $b(P)\leqslant n\beta$ 为闭区域,其上连续函数必有极值.

(2) $C_n(\beta)$ 的定义域为 $\beta\geqslant\beta_{\min}$,这里

$$\beta_{\min} \xlongequal{\text{def}} \min_{x\in\mathscr{X}} b(x). \tag{3-19}$$

信道的价值容量函数 $C(\beta)$ 定义为

$$C(\beta) \xlongequal{\text{def}} \sup_{n\geq 1} \frac{1}{n} C_n(\beta). \tag{3-20}$$

$C(\beta)$ 纯属信道的特征数，与入口分布无关；它的信息含义是，在信道入口符号的平均价值不超过 β 的情况下，每个信道符号可靠传递的最大信息量.

价值容量函数具下列性质：

1° $C_n(\beta)$ 是 $\beta \geqslant \beta_{\min}$ 的上凸连续函数.

2° 对于无记忆信道，有

(i) $C_n(\beta) = nC_1(\beta) \quad (\beta \geqslant \beta_{\min}, n = 1,2,\cdots)$；

(ii) $C(\beta) = C_1(\beta)$；

(iii) $C(\beta)$ 在 $\beta_{\min} \leqslant \beta \leqslant \beta_{\max}$ 中严格单增，且

$$C(\beta) = \max_{P:b(P)=\beta} = I(P;Q), \tag{3-21}$$

其中

$$\beta_{\max} \xlongequal{\text{def}} \min[b(P):I(P;Q) = C].$$

这里 C 是无记忆信道容量，有

$$C = \max_{\beta\geq\beta_{\min}} C(\beta) = C(\beta_{\max}). \tag{3-22}$$

3.3.2 具价值信道编码定理

考虑 k 到 n 定长分组码 (f,g)，

$$f:\mathscr{U}^k \to \mathscr{X}^n, \quad g:\mathscr{Y}^n \to \mathscr{V}^k = \mathscr{U}^k,$$

假定翻码 f 是 $\mathscr{U}^k$ 到 $\mathscr{X}^n$ 的某个子集上的一一映射，这个子集记之为 $\mathscr{C}_M$，称之为码字集合（简称为信道码），写为

$$\begin{aligned}\mathscr{C}_M \xlongequal{\text{def}} f(\mathscr{U}^k) &= \{x = f(u), u \in \mathscr{U}^k\} \\ &= \{x_1, x_2, \cdots, x_M\} \subset \mathscr{X}^n;\end{aligned}$$

这里每个映像 x_m，称为一个码字；它的逆像是一个信源消息 $u_m = f^{-1}(x_m) \in \mathscr{U}^k$；码字总数为 M 个，$M = |f(\mathscr{U}^k)| = |\mathscr{C}_M|$，且称

$$R_C \xlongequal{\text{def}} \frac{\log M}{n} \tag{3-23}$$

为 n 长信道码的传输速率.

设信源的全体 k 长消息为

$$\mathscr{U}^k = [u_1, u_2, \cdots, u_M] \quad (M = |\mathscr{U}|^k),$$

它们与 n 长信道码字集合 $\mathscr{C}_M = \{x_1, x_2, \cdots, x_M\}$ 之间的一一对应关系形成了翻码 f，

$$f:\mathscr{U}^k \leftrightarrow \mathscr{C}_M, \quad x_m = f(u_m) \quad (m = 1,2,\cdots,M).$$

此时，译码 g 实际是 $\mathscr{Y}^n$ 到 $\mathscr{C}_M \cup [?]$ 的一种映射，

$$g:\mathscr{Y}^n \to \mathscr{C}_M \cup [?],$$

这里符号“?”表示检错,按误差计算;从码字集合 $\mathscr{C}_M$ 再到 $\mathscr{X}^k \doteq \mathscr{U}^k$ 的译码则按映射 f 的逆像判定.因此,发送消息 $u_m \in \mathscr{U}^k$ 时的传输误差概率为

$$p_{e,m} \xlongequal{\text{def}} \sum_{y:g(y)\neq x_m} Q(y/x_m) \quad (m = 1,2,\cdots,M). \tag{3-24}$$

现在的问题是,对于给定信道,是否存在这样的“好码”,使得其中码字个数足够多,以便一一对应地传递被压缩之后的必用消息,同时还要使误差概率 $p_{e,m}$ 一致地小,或者说,在传输速率尽可能大的同时,保证尽量小的误差概率?

定理 2 (具价值信道编码定理) 设无记忆信道$[\mathscr{X};Q(y/x),\mathscr{Y}]$的价值函数为 $b(x), x \in \mathscr{X}$,价值容量函数为 $C(\beta)$,则对于任意 $\beta \geqslant \beta_{\min}, \rho > 0, \sigma > 0$ 及 $\epsilon > 0$,存在 n 长信道码 $\mathscr{C}_M = \{x_1, x_2, \cdots, x_M\} \subset \mathscr{X}^n$ 及译码 $g:\mathscr{Y}^n \to \mathscr{C}_M \cup [?]$,满足

1° $b(x_m) \leqslant n(\beta + \rho), p_{e,m} < \epsilon \quad (m = 1,2,\cdots,M)$;

2° $M \geqslant 2^{n(C(\beta)-\sigma)}$.

概括地说,存在这样的 n 长码,每个码符平均价值不超过 β,使用每一个码字的误差概率不超过 ϵ,而且码字总数 M 接近 $2^{nC(\beta)}$.或者说,理论上存在这样的“好码”,使每个码字实用误差概率小于 ϵ,每个码符耗费不超过 β,而且传输速率接近信道容量 $C(\beta)$.

推论 1 (无记忆信道编码定理) 设无记忆信道$[\mathscr{X};Q(y/x),\mathscr{Y}]$的容量为 C,则对于任一 $R < C$,及 $\epsilon > 0$,存在 n 长信道码 $\mathscr{C}_M = \{x_1, \cdots, x_M\} \subset \mathscr{X}^n$ 及译码 $g:\mathscr{Y}^n \to \mathscr{C}_M \cup [?]$,使

$$p_{e,m} < \epsilon \quad (m = 1,2,\cdots,M); \quad M \geqslant 2^{Rn}.$$

3.4 误差概率指数界

3.4.1 误差概率的指数形式

信道编码定理 2 及其推论 1 表明,对于无记忆信道,当其传输速率小于信道容量时,必存在 n 长信道码及适当译码,使每个码字的误差概率均可小于任何预定正数.但应注意,这里使误差概率小的关键是增大码长 n,而码长增大必定会给选码及实现通信造成困难.因此,希望在满足误差限度的情况下,实用码长能尽量地短.理论上解决这一问题时可用如下定理.

定理 3 设无记忆信道为$[\mathscr{X};Q(y/x),\mathscr{Y}]$,则对于任意正整数 $n, R > 0$ 及 $0 \leqslant \lambda \leqslant 1$,存在具有 M 个码字的 n 长信道码 $\mathscr{C}_M = \{x_1, x_2, \cdots, x_M\} \subset \mathscr{X}^n$ 及译码 $g:\mathscr{Y}^n \to \mathscr{C}_M$,满足

1° $p_{e,m} < 4\exp[-nE(R)] \quad (1 \leqslant m \leqslant M)$;

2° $M \geqslant \exp[nR]$.

其中

$$p_{e,m} \xlongequal{\text{def}} \sum_{y:g(y)\neq x_m} Q(y/x_m);$$

$$E(R) \xlongequal{\text{def}} \max_{0\leqslant\lambda\leqslant 1}\{\max_P(E(\lambda,P)-\lambda R)\} \quad (R>0); \tag{3-25}$$

$$E(\lambda,P) \xlongequal{\text{def}} -\ln\sum_{y\in\mathscr{Y}}[\sum_{x\in\mathscr{X}} p(x)Q(y/x)^{\frac{1}{1+\lambda}}]^{1+\lambda} \quad (0\leqslant\lambda\leqslant 1),$$

其中 $p(x), x\in\mathscr{X}$,为任一信道入口分布.最大值是对所有入口分布 P 而取的.

3.4.2 指数界信道编码定理

定理3只是给出误差概率 $p_{e,m}$ 的指数形式,关键是"随机编码指数" $E(R)$ 的性能问题,它纯属信道转移概率 Q 的特征.取什么样的信道和哪些 R 值才能使 $E(R) > 0$ 以及作为 R 的函数,$E(R)$ 具有怎样的特性?见以下论述.

定理4 (指数界信道编码定理) 设无记忆信道$[\mathscr{X}, Q, \mathscr{Y}]$的容量 $C>0$,则当 $0\leqslant R<C$ 时,随机编码指数 $E(R)$ 是 R 的单减下凸($\cup$)正值函数,即 $0<E(R)$,$0\leqslant R<C$.

定理3证明的难度相当大,详见参考文献[3].下面例2说明 $E(R)$ 的计算也很繁难.

例2 设二元对称信道为$[\mathscr{X}, Q(y/x), \mathscr{Y}]$,其中 $\mathscr{X}=\{0,1\}=\mathscr{Y}$,$Q(0/0)=1-\epsilon=Q(1/1)$,

$$Q(1/0)=Q(0/1)=\epsilon,\quad 0<\epsilon<\frac{1}{2}.$$

求该信道的随机编码指数 $E(R)$.

解 按定义(3-25)式,有

$$E(R) \xlongequal{\text{def}} \max_{0\leqslant\lambda\leqslant 1}\max_P g(\lambda;P,R),$$

$$g(\lambda;P,R)=E(\lambda,P)-\lambda R,$$

$$E(\lambda,P) \xlongequal{\text{def}} -\ln\sum_{y=0}^{1}[\sum_{x=0}^{1} p(x)Q(y/x)^{\frac{1}{1+\lambda}}]^{1+\lambda}.$$

若记入口分布 $p(0)=p$,$p(1)=1-p$,则上面诸式中的二元参量 $P=[p,(1-p)]$ 可视为单个变量 p 的函数,$0\leqslant p\leqslant 1$.此时 $E(\lambda,p)$ 可写成

$$E(\lambda,p)=-\ln\{[p(1-\epsilon)^{\frac{1}{1+\lambda}}+(1-p)\epsilon^{\frac{1}{1+\lambda}}]^{1+\lambda}+ [p\epsilon^{\frac{1}{1+\lambda}}+(1-p)(1-\epsilon)^{\frac{1}{1+\lambda}}]^{1+\lambda}.$$

由对称性可知,

$$\max_{0\leqslant p\leqslant 1} E(\lambda,p)=E(\lambda,\frac{1}{2}),$$

从而有

$$\max_{0\leqslant p\leqslant 1} g(\lambda;p,R)=g(\lambda;\frac{1}{2},R)=E(\lambda,\frac{1}{2})-\lambda R,$$

$$E(R)=\max_{0\leqslant\lambda\leqslant 1} g(\lambda;\frac{1}{2},R).$$

经过复杂的计算,运用一定的分析技巧(详见参考文献[3])最终得到

$$E(R) = \begin{cases} \ln 2 - 2\ln(\sqrt{\varepsilon} + \sqrt{1-\varepsilon}) - R & (\text{当 } 0 \leqslant R \leqslant R_1); \\ T_\varepsilon(\delta) - H(\delta) & (\text{当 } R_1 \leqslant R \leqslant R). \end{cases} \tag{3-26}$$

其中

$$R_1 = \ln 2 - H\left(\frac{\sqrt{\varepsilon}}{\sqrt{\varepsilon} + \sqrt{1-\varepsilon}}\right),$$

$$T_\varepsilon(\delta) = -\delta\ln\varepsilon - (1-\delta)\ln(1-\varepsilon),$$

$$R_0 = \ln 2 - H(\varepsilon) = C, \quad R = \ln 2 - H(\delta).$$

参 考 文 献

1 Gallager R G. Information theory and reliable communication. New York: Wiley, 1968.

2 McEliece R J. The theory of information and coding. Encyclopedia of Math and its Applications, vol 3. Reading, Mass: Addison-Wesley, 1977.

3 孟庆生著.信息论.西安:西安交通大学出版社,1986.

·经济数学卷·

第19篇

人工神经网络

编　者　杨行峻　郑君里
审校者　赵明生

目　录

引　言

神经网络是由非常大量的**神经元**构成的系统，每个神经元是一个简单的非线性处理单元．人工神经网络由人工神经元构成，它在不同层次和方面模仿人脑神经系统的信息存储、检索及处理功能，诸如大规模并行处理、分布式存储、学习能力和自适应性等．

D.E.Rumelhart 和 J.L.McClelland 关于并行分布处理系统的经典工作（见文献[3]）和 Hopfield的研究工作（见文献[4]）自 20 世纪 80 年代中期出现以来，极大地推动了人工神经网络研究工作的开展，成就很大．同时它也是一门正在蓬勃发展中的新学科，很多理论问题有待于解决，特别需加强与模糊、混沌、遗传算法相结合．

神经网络的应用几乎遍及所有自然科学、工业和经济学领域．用它来解决优化、联想、模式识别、函数逼近、时间序列预测、非线性系统辨识和控制、参数提取、故障诊断、信号检测和编码、人工智能系统等许多问题时，都取得了其它很多方法难以达到的成果．

神经网络有多种模型．按照学习算法可分为有监督的学习和无监督（自组织）的学习两大类．本篇介绍使用最普遍的三种模型：多层前向神经网络（MLFN）、Hopfield 神经网络和自组织特征映射（SOFM）神经网络；第一种采用有监督的学习算法，第三种采用自组织学习算法．

1　多层前向神经网络(MLFN)

1.1　MLFN 的基本原理

1.1.1　MLFN 的结构

MLFN 由一个输入层（层号 $l=0$）、$L-1$ 个隐层（层号 $l=(1\sim L-1)$）和一个输出层（层号 $l=L$）构成，第 l 层包含 N_l 个神经元，其输出记为 $x_i^{(l)}, i=1,2,\cdots,N_l$．$X^{(0)}=(x_1^{(0)},x_2^{(0)},\cdots,x_{N_0}^{(0)})$ 是网络的输入矢量，$X^{(0)}\in \mathbf{R}^{N_0}$．$X^{(L)}=(x_1^{(L)},x_2^{(L)},\cdots,x_{N_L}^{(L)})$ 是网络的输出矢量，$X^{(L)}\in \mathbf{R}^{N_L}$．网络结构如图 1-1 所示，信号由输入单方向逐层送往输出，层内各神经元互不传送信号．若 $N_L=1$ 且 $N_0>1$，则称其为 **MISO L 层网络**；若 $N_L>1$ 且 $N_0>1$，则称其为 **MIMO L 层网络**．设 $X^{(l)}=(x_1^{(l)},x_2^{(l)},\cdots,x_{N_l}^{(l)})$，则由 $X^{(l-1)}$ 求 $X^{(l)}$ 的计算公式如下：

$$x_i^{(l)}=f_l(I_i^{(l)}),\quad i=1,2,\cdots,N_l. \tag{1-1}$$

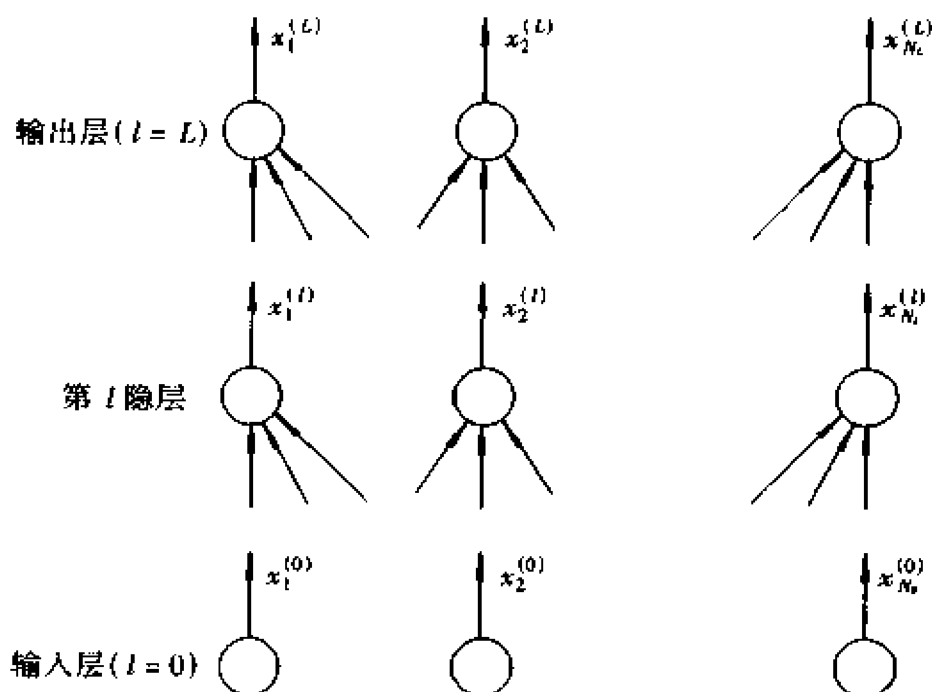


图 1-1

$$I_i^{(l)} = \varphi_l(X^{(l-1)}, W_i^{(l)}, \theta_i^{(l)}), \tag{1-2}$$

其中 $f_l(\cdot)$ 是 l 层的神经元函数，$I_i^{(l)}$ 是 l 层第 i 神经元的净输入，$W_i^{(l)}$ 和 $\theta_i^{(l)}$ 分别为 l 层第 i 神经元的权矢量和阈值参数；$W_i^{(l)} = (w_{i1}^{(l)}, w_{i2}^{(l)}, \cdots, w_{iN_{l-1}}^{(l)})$，$W_i^{(l)} \in \mathbf{R}^{N_{l-1}}$。这样，当使(1-1)式和(1-2)式中诸函数得到定义且诸参数确定后，即可由任何输入矢量 $X^{(0)}$ 逐层上推，直至求得网络输出矢量 $X^{(L)}$。这称为前向运算。

1.1.2　常用的神经元函数 $f_l(\cdot)$ 和 $\varphi_l(\cdot)$

1. 线性函数

$$f_l(I_i^{(l)}) = I_i^{(l)}, \tag{1-3}$$

$$\varphi_l(X^{(l-1)}, W_i^{(l)}, \theta_i^{(l)}) = X^{(l-1)} \cdot W_i^{(l)} + \theta_i^{(l)}. \tag{1-4}$$

2. 指示函数(硬限幅函数)

$$f_l(I_i^{(l)}) = \operatorname{sgn}(I_i^{(l)}) = \begin{cases} 1, & I_i^{(l)} \geqslant 0, \\ 0, & I_i^{(l)} < 0. \end{cases} \tag{1-5}$$

$\varphi_l(\cdot)$ 由(1-4)式计算.

3. Sigmoid 函数

$$f_l(I_i^{(l)}) = (1 + \mathrm{e}^{-I_i^{(l)}})^{-1} \tag{1-6}$$

或

$$f_l(I_i^{(l)}) = \tanh(I_i^{(l)}), \tag{1-7}$$

$\varphi_l(\cdot)$ 由(1-4)式计算.

4. 径向基函数(RBF)

$$f_l(I_i^{(l)}) = (\sqrt{2\pi}\sigma_i^{(l)})^{-1}e^{-I_i^{(l)}}, \quad \sigma_i^{(l)} > 0, \tag{1-8}$$

$$I_i^{(l)} = \varphi_l(X^{(l-1)}, W_i^{(l)}) = \frac{1}{2}\| X^{(l-1)} - W_i^{(l)} \|^2/(\sigma_i^{(l)})^2, \tag{1-9}$$

或

$$f_l(I_i^{(l)}) = (2\pi|\Sigma_i^{(l)}|)^{-\frac{1}{2}}e^{-I_i^{(l)}}, \tag{1-10}$$

$$I_i^{(l)} = \varphi_l(X^{(l-1)}, W_i^{(l)}) = \frac{1}{2}(X^{(l-1)} - W_i^{(l)})(\Sigma^{(l)})_i^{-1}(X^{(l-1)} - W_i^{(l)})^{\mathrm{T}}, \tag{1-11}$$

其中 $\Sigma_i^{(l)}$ 是一个 $N_{l-1} \times N_{l-1}$ 对称正定矩阵.

1.1.3　损失函数和风险函数

为简化又不失一般性,下面只讨论MISO网络,另外,将网络的输入和输出特别表示为 $X = X^{(0)}, X = (x_1, x_2, \cdots, x_N), N = N_0; y = x_1^{(L)}$.则 MLFN 所完成的由输入至输出映射关系可以表示为 $y = f(X, \xi), \xi \in \Lambda$,其中 ξ 表示网络中所有的权和阈值参数,Λ 表示其定义域.更一般而言,Λ 可表示网络结构(层数和每层的神经元数).

假设需要由网络实现的映射用某特定输入 X 产生某特定$\hat{y}$的条件概率密度函数 $p(\hat{y}/X)$ 描述,其特例是二者之间有固定函数关系:$\hat{y} = \hat{f}(X)$.网络的理想输出$\hat{y}$与实际输出 y 之间的差异可以用**损失函数**$L(\hat{y}, y)$ 描述,也可以表示为 $L(\hat{y}, f(X, \xi))$.常用的损失函数定义有以下几种.

1. 模式识别(分类问题)

$\hat{y}$只能取值为0或1,分别表示 X 属于两种类别中的某一类;相应地 $y = f(X, \xi)$ 也只能取0或1值.其损失函数定义如下:

$$L(\hat{y}, f(X, \xi)) = \begin{cases} 0, & \hat{y} = f(X, \xi), \\ 1, & \hat{y} \neq f(X, \xi). \end{cases} \tag{1-12}$$

2. 回归估计(联想问题)

$\hat{y}$和 y 皆取实数值,其损失函数定义为

$$L(\hat{y}, f(X, \xi)) = (\hat{y} - f(X, \xi))^2. \tag{1-13}$$

损失函数对于$\hat{y}$和 X 的全集合统计平均值定义为**风险函数**$R(\xi)$.设$\hat{y}$和 X 的联合概率分布函数是 $P(\hat{y}, X)$,则 $R(\xi)$ 可以表示为

$$R(\xi) = \int L(\hat{y}, f(X, \xi))\mathrm{d}P(\hat{y}, X). \tag{1-14}$$

最小风险 $R(\xi_0)$ 是 $R(\xi)$ 的下界,可表示为

$$R(\xi_0) = \inf_{\xi \in \Lambda} \{R(\xi)\}. \quad (1\text{-}15)$$

1.1.4 学习过程的经验风险最小归纳原理

求一个 MLFN 的最佳参数 ξ_0 使 $R(\xi)$ 达到最小,称为一个**学习过程**.一般情况下 $P(\hat{y}, X)$ 对于一个特定的问题是确定且未知的,这时 ξ_0 难以求得.可行的方法是由一个监督器(或称教师)给出 M 组训练样本 $(\hat{y}_1, X_1), (\hat{y}_2, X_2), \cdots, (\hat{y}_M, X_M)$,构成一个训练集合,通过对此集合的学习来找到最佳参数.为此首先构造**经验风险函数**(**或称代价函数**) $R_e(\xi)$ 如下:

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M} L(\hat{y}_m, f(X_m, \xi)). \quad (1\text{-}16)$$

设 ξ_a 为使 $R_e(\xi)$ 达到最小值的参数,即

$$R_e(\xi_a) = \inf_{\xi \in \Lambda} \{R_e(\xi)\}, \quad (1\text{-}17)$$

可以将 ξ_a 作为 MLFN 的最佳参数.这一学习策略称为**经验风险最小归纳原理**(**ERM**).

一个基于 ERM 的学习过程涉及下列问题:

(1) 如何求 ξ_a?

(2) 虽然 ξ_a 使 $R_e(\xi)$ 达到最小,$R(\xi_a)$ 是否足够小?$R(\xi_a)$ 与 $R(\xi_0)$ 之间差距有多大?

(3) 在训练集规模 M 一定的条件下,为了使 $R(\xi_a)$ 达到最小值(即网络具有良好推广性能),应如何选择网络规模(层数和每层的神经元个数)?

以下诸节将陆续讨论这些问题.

1.2 MLFN 的参数学习算法(求 ξ_a)

常用的 MLFN 有两种.第一种的隐层神经元取 Sigmoid 函数,输出层神经元取线性函数或 Sigmoid 函数,这种网络常称为 MLP.第二种网络的隐层(只有一个隐层)神经元取径向基函数,输出层神经元取线性函数,这种网络常称为 RBF.下面分别讨论这两种网络的学习算法.

1.2.1 MLP 的参数学习算法

1. 离线批处理 BP 算法

对于 MISO L 层 MLP,按(1-13)式计算损失函数时经验风险函数的计算公式如下:

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - f(X_m, \xi))^2 = \frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - x_{m1}^{(L)})^2. \quad (1\text{-}18)$$

为求得使 $R_e(\xi)$ 达到极小值的最佳参数矢量 ξ_a,可以采用最陡下降算法.首先设置随机初值 $\xi(0)$,再按节拍 $k = 0, 1, 2, \cdots$ 进行递推计算

$$\xi(k+1)=\xi(k)+\Delta\xi(k),$$
$$\Delta\xi(k)=-\alpha\nabla_{\xi}R_e(\xi)\big|_{\xi=\xi(k)}, \tag{1-19}$$

其中 $\alpha>0$,称为步幅.若 α 取较小值,使得 $\|\Delta\xi(k)\|\ll 1$($\|\cdot\|$ 表示一个矢量的模值),下列公式成立:

$$R_e(\xi(k+1))=R_e(\xi(k))+\Delta R_e(\xi(k)),$$
$$\Delta R_e(\xi(k))\approx\Delta\xi(k)\cdot\nabla_{\xi}R_e(\xi)\big|_{\xi=\xi(k)}=-\alpha\left\|\nabla_{\xi}R_e(\xi)\big|_{\xi=\xi(k)}\right\|^2\leqslant 0.$$

这表明,只要 α 足够小,$R_e(\xi(k))$ 将随着 k 的增加而下降,因此 $\xi(k)$ 将趋向于 $R_e(\xi)$ 的一个局部极小点.若此点的 $R_e(\xi)$ 值足够小,即可以作为所需的解 ξ_a.

在递推计算中,(1-19) 式可表示为各个权和阈值的递推计算(二者类似,只列前者,下文同此).

$$w_{ij}^{(l)}(k+1)=w_{ij}^{(l)}(k)+\Delta w_{ij}^{(l)}(k),\quad k=0,1,2,\cdots,$$
$$\Delta w_{ij}^{(l)}(k)=-\alpha\frac{\partial R_e(\xi)}{\partial w_{ij}^{(l)}}\bigg|_{\xi=\xi(k)}. \tag{1-20}$$

这样,递推计算归结为求上式右端的偏微分.引用(1-18) 式,得到

$$\frac{\partial R_e(\xi)}{\partial w_{ij}^{(l)}}\bigg|_{\xi=\xi(k)}=\frac{-2}{M}\sum_{m=1}^{M}\left[(\hat{y}_m-x_{m1}^{(L)})\frac{\partial x_{m1}^{(L)}}{\partial w_{ij}^{(l)}}\right]\bigg|_{\xi=\xi(k)}, \tag{1-21}$$

其中 $x_{m1}^{(L)}$ 由前向计算求得,$\hat{y}_m$ 在训练集中给定.现在问题归结为求下列偏微分

$$\frac{\partial x_{mq}^{(l')}}{\partial w_{ij}^{(l)}}\bigg|_{\xi=\xi(k)},\quad q=1\sim N_{l'},\, l\leqslant l',\, l'=1\sim L.$$

只要能够计算它,便能计算(1-21) 式右侧的偏微分.引用(1-1) 至(1-4) 式和(1-6) 式,得到

$$x_{mq}^{(l')}=f_{l'}(I_{mq}^{(l')})=f_{l'}\Big(\sum_{p=1}^{N_{l'-1}}w_{qp}^{(l')}x_{mp}^{(l'-1)}+\theta_q^{(l')}\Big), \tag{1-22}$$

$$\frac{\partial x_{mq}^{(l')}}{\partial w_{ij}^{(l)}}=\frac{\mathrm{d}f_{l'}(I_{mq}^{(l')})}{\mathrm{d}I_{mq}^{(l')}}\cdot\frac{\partial I_{mq}^{(l')}}{\partial w_{ij}^{(l)}}. \tag{1-23}$$

(1-23) 式右侧的第 2 项(偏微分项) 可以分以下两种情况来计算:

(1) $l=l'$,

$$\frac{\partial I_{mq}^{(l')}}{\partial w_{ij}^{(l)}}=\begin{cases}x_j^{(l'-1)}, & q=i,\\ 0, & q\neq i.\end{cases} \tag{1-24}$$

(2) $l<l'$,

$$\frac{\partial I_{mq}^{(l')}}{\partial w_{ij}^{(l)}}=\sum_{p=1}^{N_{l'-1}}w_{qp}^{(l')}\frac{\partial x_{mp}^{(l'-1)}}{\partial w_{ij}^{(l)}}. \tag{1-25}$$

以下又分成两种情况:

1) 若 $l'-1=l$,则(1-25) 式右侧的各项偏微分可以用(1-23) 和(1-24) 式求得.

2) 若 $l'-1>l$,则式(1-25) 右侧各偏微分可以用(1-23) 式和再次使用(1-25)

式求得.

(1-23)式右侧第1项(微分项)可分下列两种情况进行计算.

1) $f_r(\cdot)$ 为线性函数((1-3)式),

$$\frac{\mathrm{d}f_r(I_{mq}^{(r)})}{\mathrm{d}I_{mq}^{(r)}} = 1. \tag{1-26}$$

2) $f_r(\cdot)$ 为第一种Sigmoid函数((1-6)式),

$$\frac{\mathrm{d}f_r(I_{mq}^{(r)})}{\mathrm{d}I_{mq}^{(r)}} = f_r(I_{mq}^{(r)})(1 - f_r(I_{mq}^{(r)})) = x_{mq}^{(r)}(1 - x_{mq}^{(r)}). \tag{1-27}$$

上述全部计算可以规整为下列便于进行编程计算的形式.由(1-20)和(1-21)式,每一节拍 k 的权调整量可用下式计算:

$$\Delta w_{ij}^{(l)}(k) = \frac{2\alpha}{M}\sum_{m=1}^{M}(\hat{y}_m - x_{m1}^{(L)}(k))\frac{\partial x_{m1}^{(L)}}{\partial w_{ij}^{(l)}}\bigg|_{\xi=\xi(k)}. \tag{1-28}$$

令 $\delta_{mi}^{(l)}$ 表示 $x_{m1}^{(L)}$ 对 $I_{mi}^{(l)}$ 的偏导数,再利用(1-2)和(1-4)式,得到

$$\frac{\partial x_{m1}^{(L)}}{\partial w_{ij}^{(l)}} = \frac{\partial x_{m1}^{(L)}}{\partial I_{mi}^{(l)}}\cdot\frac{\partial I_{mi}^{(l)}}{\partial w_{ij}^{(l)}} = \delta_{mi}^{(l)}\cdot x_{mj}^{(l-1)}. \tag{1-29}$$

这样,(1-28)式可改写为下列形式:

$$\Delta w_{ij}^{(l)}(k) = \frac{2\alpha}{M}\sum_{m=1}^{M}(\hat{y}_m - x_{m1}^{(L)}(k))\delta_{mi}^{(l)}(k)x_{mj}^{(l-1)}(k). \tag{1-30}$$

$\delta_{mi}^{(l)}(k)$ 分下列两种情况进行计算:

(1) $l = L$,只需计算 $\delta_{m1}^{(L)}(k)$.注意到 $x_{m1}^{(L)}(k) = f_L(I_{m1}^{(L)}(k))$,由(1-26)和(1-27)式可得

$$\delta_{m1}^{(L)}(k) = \begin{cases}1, f_L(\cdot)\text{ 为线性函数},\\ x_{m1}^{(L)}(k)(1 - x_{m1}^{(L)}(k)), f_L(\cdot)\text{ 为第一种 Sigmoid 函数}.\end{cases} \tag{1-31}$$

(2) $l = 1,2,\cdots,L-1$,按以下推导由 $\delta_{mp}^{(l+1)}(k)$, $p = 1,2,\cdots,N_{l+1}$,求 $\delta_{mi}^{(l)}(k)$, $i = 1,2,\cdots,N_l$:

$$\delta_{mi}^{(l)}(k) = \frac{\partial x_{m1}^{(L)}(k)}{\partial I_{mi}^{(l)}(k)} = \sum_{p=1}^{N_{l+1}}\frac{\partial x_{m1}^{(L)}(k)'}{\partial I_{mp}^{(l+1)}(k)}\cdot\frac{\partial I_{mp}^{(l+1)}(k)}{\partial I_{mi}^{(l)}(k)},$$

$$\frac{\partial I_{mp}^{(l+1)}(k)}{\partial I_{mi}^{(l)}(k)} = \frac{\partial I_{mp}^{(l+1)}(k)}{\partial x_{mi}^{(l)}(k)}\cdot\frac{\partial x_{mi}^{(l)}(k)}{\partial I_{mi}^{(l)}(k)}$$

$$= w_{pi}^{(l+1)}(k)x_{mi}^{(l)}(k)(1 - x_{mi}^{(l)}(k)),$$

$$\frac{\partial x_{m1}^{(L)}(k)}{\partial I_{mp}^{(l+1)}(k)} = \delta_{mp}^{(l+1)}(k),$$

$$\delta_{mi}^{(l)}(k) = \sum_{p=1}^{N_{l+1}}\delta_{mp}^{(l+1)}(k)w_{pi}^{(l+1)}(k)x_{mi}^{(l)}(k)(1 - x_{mi}^{(l)}(k)), \quad i = 1 \sim N_l. \tag{1-32}$$

利用(1-32)式,可由 $\delta_{m1}^{(L)}(k)$ 求得 $\delta_{mi}^{(L-1)}(k)$, $i = 1,2,\cdots,N_{L-1}$,再用后者求得

$\delta_{mi}^{(L-2)}(k), i = 1,2,\cdots,N_{L-2}$;这样由上至下直至求得 $\delta_{mi}^{(1)}(k), i = 1,2,\cdots,N_1$.

至此 $\Delta w_{ij}^{(l)}(k)$ 的计算已解决.由于 $\delta_{mi}^{(l)}(k)$ 的计算是由输出层($l = L$)反推至输入端的第一隐层($l = 1$),所以此学习算法称为 **BP**(BP 是 Back Propagation 的缩写)**算法**.由于学习前已采集到全部训练数据,学习可以脱离现场进行,且每一步权调整递推计算都需要调用训练集中的全部数据,所以称之为**离线批处理 BP 算法**.阈值参数 $\theta_i^{(l)}$ 的递推计算方法与权参数类似(只需将 $\theta_i^{(l)}$ 看成第 $l-1$ 层一个输出恒等于 1 的神经元至第 l 层第 i 神经元的信号传送权值即可),此处不再赘述.

2. 离线随机梯度 BP 算法

与批处理算法的区别是每个权调整节拍 k 只从采集好的训练集中随机或依次地取出一组训练数据,并且记之为 $(\hat{y}(k), \boldsymbol{X}(k))$.将 $\boldsymbol{X}(k)$ 输入网络,通过前向计算求得 $x_1^{(L)}(k)$ 和网络各层输出 $x_i^{(l)}(k)$.递推计算从随机设置的各个权初值 $w_{ij}^{(l)}(0)$ 出发,按下式进行计算:

$$
\begin{aligned}
w_{ij}^{(l)}(k+1) &= w_{ij}^{(l)}(k) + \Delta w_{ij}^{(l)}(k), \quad k = 0,1,2,\cdots, \\
\Delta w_{ij}^{(l)}(k) &= 2\alpha(k)[\hat{y}(k) - x_1^{(L)}(k)] \left.\frac{\partial x_1^{(L)}(k)}{\partial w_{ij}^{(l)}}\right|_{\xi=\xi(k)}.
\end{aligned} \tag{1-33}
$$

此式右侧的偏微分仍可采用(1-29)至(1-32)式给出的公式来计算.$\alpha(k) > 0$ 是一个随 k 而变化的步幅函数.当(1-34)式条件成立时,随着 k 的增加,$\boldsymbol{\xi}(k)$ 将收敛到 $R_e(\boldsymbol{\xi})$ 的一个局部极小点,

$$\sum_{k=0}^{\infty}\alpha(k) = \infty, \sum_{k=0}^{\infty}\alpha^2(k) < \infty. \tag{1-34}$$

当 $k = 0$ 时,$\alpha(k) = \alpha_0 > 0$;当 $k \geqslant 1$ 时,$\alpha(k) = \alpha_0/k$.这是满足上述条件的一种步幅函数.

由于此算法的每步权调整递推计算只调用训练集中的一组数据,所以就单步而言权的调整具有随机性,这是随机梯度算法名称的由来.模拟实验结果表明,在一些实际情况中,随机梯度算法的计算开销较小且收敛极小点的质量较高.还有一种折衷方案是每递推一步调用一小批(例如 10 组)训练样本.

3. 在线算法

在线算法与离线算法的区别在于后者是在训练数据采集完成后离开现场对网络进行训练,而前者是实时实地进行数据采集和网络训练,从而使网络性能自适应于环境的变化.设按照实时 n 采集的训练数据是 $(\hat{y}(n), \boldsymbol{X}(n))$,那么随 n 而改变的实时经验风险函数定义如下:

$$R_e(\boldsymbol{\xi}(n)) = \frac{1}{\mathcal{N}}\sum_{\gamma=n-\mathcal{N}}^{n-1}(\hat{y}(\gamma) - x_1^{(L)}(\gamma))^2,$$

其中 $x_1^{(L)}(\gamma)$ 是输入为 $\boldsymbol{X}(\gamma)$ 时网络的输出.$\mathcal{N}$ 是 n 时刻前取平均的样本数.$\mathcal{N}$ 值越大则学习的统计效果越好而自适应性变弱.网络的学习算法与前述离线情况并无区别.

1.2.2　MLP参数学习算法的改进

1. 惯性学习算法

在标准BP算法中，步幅 α 是一很难选择的参数. 若 α 选得太小，则收敛速度太慢；若 α 选得太大，则会在极小点附近产生振荡以致不能收敛. 如果将(1-21)式的偏导数用 $S_{ij}^{(l)}(k)$ 来表示，则可以采用下列**惯性学习算法**：

$$\Delta w_{ij}^{(l)}(k) = \alpha_{ij}^{(l)}(k)S_{ij}^{(l)}(k) + \eta_{ij}^{(l)}(k)\Delta w_{ij}^{(l)}(k-1), \quad k = 0,1,\cdots, \tag{1-35}$$

且假设 $\Delta w_{ij}^{(l)}(-1) = 0$. 此式右侧第一项是标准学习项，右侧第二项是惯性学习项，$\alpha_{ij}^{(l)}(k)$ 和 $\eta_{ij}^{(l)}(k)$ 分别为步幅和惯性系数，应满足条件：

$$\alpha_{ij}^{(l)}(k) > 0, 1 > \eta_{ij}^{(l)}(k) > 0. \tag{1-36}$$

一种最简单的方案是对任何 $l、i、j、k$ 都取其为常数 α_0 和 η_0(例如，$\alpha_0 = 0.1$，$\eta_0 = 0.7$). 更好的方案是令其随 $l、i、j、k$ 而变化. 下面列出的几条规则给出了这种方案的一个例子.

(1) 设定初值. 对任何 $l、i、j$，$\alpha_{ij}^{(l)}(0) = \alpha_{ij}^{(l)}(1) = 0.2$. $\eta_{ij}^{(l)}(0) = \eta_{ij}^{(l)}(1) = 0.7$.

(2) 当 $k \geqslant 2$ 时，若 $S_{ij}^{(l)}(k)$、$S_{ij}^{(l)}(k-1)$、$S_{ij}^{(l)}(k-2)$ 具有相同符号，令 $\alpha_{ij}^{(l)}(k) = \mu\alpha_{ij}^{(l)}(k-1)$，$\eta_{ij}^{(l)}(k) = \mu\eta_{ij}^{(l)}(k-1)$ (例如 $\mu = 1.1$). 若 $S_{ij}^{(l)}(k)$、$S_{ij}^{(l)}(k-1)$、$S_{ij}^{(l)}(k-2)$ 具有不同符号，令 $\alpha_{ij}^{(l)}(k) = \lambda\alpha_{ij}^{(l)}(k-1)$，$\eta_{ij}^{(l)}(k) = \lambda\eta_{ij}^{(l)}(k-1)$(例如 $\lambda = 0.5$).

(3) 设定各 $\alpha_{ij}^{(l)}(k)$ 的最大值为 $\alpha_{\max} = 0.5$ 和最小值 $\alpha_{\min} = 0.01$，各 $\eta_{ij}^{(l)}(k)$ 的最大值为 $\eta_{\max} = 0.9$ 和最小值为 $\eta_{\min} = 0.1$. 如超出此范围则停止增减.

2. 输入矢量和输出矢量规格化

对于输入矢量 $\boldsymbol{X}$ 和输出 y 予以规格化可以提高网络的学习速度. 以 $\boldsymbol{X}$ 为例，设其第 j 分量 x_j 的取值范围是$(x_{j\min}, x_{j\max})$，规格化是通过线性变换将 x_j 转换为 x'_j，使得对于任何 j，x'_j 的取值范围为$(-1,1)$. 此线性变换可以用 $x'_j = a_jx_j + b_j$ 表示. 满足上列取值范围的转换，应满足 $-1 = a_jx_{j\min} + b_j$，$1 = a_jx_{j\max} + b_j$. 由此可以解出变换系数 a_j 和 b_j：

$$a_j = \frac{2}{x_{j\max} - x_{j\min}}, \quad b_j = -1 - a_jx_{j\min}. \tag{1-37}$$

由各 x'_j 构成的矢量 $\boldsymbol{X}'$ 将成为MLP的输入矢量.

3. 修正BP算法(MBP)

用 $\boldsymbol{S}(k)$ 表示经验风险函数的梯度，即

$$\boldsymbol{S}(k) = \boldsymbol{\nabla}_\xi R_e(\xi) \Big|_{\xi = \xi(k)}. \tag{1-38}$$

令 $\boldsymbol{d}(k)$ 为

$$\boldsymbol{d}(k) = \begin{cases} -\boldsymbol{S}(k), & k = 0; \\ -\boldsymbol{S}(k) + \beta_{k-1}\boldsymbol{d}(k-1), & k \geqslant 1, \end{cases} \tag{1-39}$$

其中　$$\beta_{k-1} = \|\boldsymbol{S}(k)\|^2 / \|\boldsymbol{S}(k-1)\|^2,$$

则参数调整公式为

$$\xi(k+1)=\xi(k)+\Delta\xi(k),$$
$$\Delta\xi(k)=\alpha^{*}d(k), \tag{1-40}$$

其中 α^{*} 是一个待定的步幅系数.设 $R_e(\xi(k+1))=R_e(\xi(k)+\alpha d(k))$,则 α^{*} 是使其达到最小的系数.再设 $R_e(\xi(k+1))$ 在 $\alpha=\alpha^{*}$ 附近可以用一个二阶多项式 $R_e(\xi(k)+\alpha d(k))=a_0+a_1\alpha+a_2\alpha^2$ 来近似,这样,α^{*} 可通过下列**线性搜索程序**求得:设 α_1 和 α_2 是 α 的两个取值.

步 1　令 $\alpha_1=0,\alpha_2=1$,令相应的 $\xi(k)+\alpha d(k)$ 为 $\xi_1(k+1)$ 和 $\xi_2(k+1)$.若 $R_e(\xi_2(k+1))<R_e(\xi_1(k+1))$,则可令 $\alpha_3=\zeta_1>1$,并求出相应的 $\xi_3(k+1)$.

(1-1)　若 $R_e(\xi_3(k+1))>R_e(\xi_2(k+1))$,则 α^{*} 必然在$(0,\zeta_1)$间隔内.这样,可以由联立方程 $R_e(\xi_i(k+1))=a_0+a_1\alpha_i+a_2\alpha_i^2, i=1,2,3$,将系数 a_0、a_1、a_2 求得.令$(a_0+a_1\alpha+a_2\alpha^2)$对 α 的导数为0,即可求得 $\alpha^{*}=-a_1/2a_2$.

(1-2)　若 $R_e(\xi_3(k+1))<R_e(\xi_2(k+1))$,则 α^{*} 必然在$(0,1)$间隔外,这时可设 $\alpha_4=\zeta_2>\zeta_1$.若 $R_e(\xi_4(k+1))>R_e(\xi_3(k+1))$,则 α^{*} 必然在$(1,\zeta_2)$间隔内,此时可以用 α_2、α_3、α_4 将 a_0、a_1、a_2 解出来.若 $R_e(\xi_4(k+1))<R_e(\xi_3(k+1))$,则可进一步设 $\alpha_5=\zeta_3>\zeta_2$,这一过程可继续下去直到确定 α^{*} 所处的间隔并将其求出.

步 2　若 $R_e(\xi_2(k+1))>R_e(\xi_1(k+1))$,则可令 $\alpha_3=\rho_1<1$,并求出相应的 $\xi_3(k+1)$.

(2-1)　若 $R_e(\xi_3(k+1))<R_e(\xi_1(k+1))$,则 α^{*} 必然在$(0,1)$间隔内.这时可以用 α_1、α_2、α_3 求出系数 a_0、a_1、a_2,从而可求得 α^{*}.

(2-2)　若 $R_e(\xi_3(k+1))>R_e(\xi_1(k+1))$,则可设 $\alpha_4=\rho_2<\rho_1$.若 $R_e(\xi_4(k+1))<R_e(\xi_1(k+1))$,则 α^{*} 必然在(o,ρ_1)之间.这时可用 α_1、α_4、α_3 解出 a_0、a_1、a_2 并求得 α^{*};反之,若 $R_e(\xi_4(k+1))>R_e(\xi_1(k+1))$,则设 $\alpha_5=\rho_3<\rho_2$,并继续这一过程直至求得 α^{*}.

步 3　检查 α^{*} 是否确实可用.

(3-1)　若 $R_e(\xi(k)+\alpha^{*}d(k))<R_e(\xi(k))$,则确定 α^{*} 为可用的.

(3-2)　反之,则在 $R_e(\xi_1(k+1)),R_e(\xi_2(k+1)),R_e(\xi_3(k+1)),\cdots$ 中找一个最小者,以相应的 $\xi_i(k+1)$ 作为待求解.

以上所述的算法也称为具有线搜索程序的共轭梯度算法.

4. 随机优化算法(RO 算法)

与其他基于最陡下降原则的优化算法一样,BP 算法或其他 BP 改进算法莫不存在着收敛于低质量局部最小点的问题.如果 $R_e(\xi)$ 的函数曲面上存在着平台(在平台上 $\nabla_\xi R_e(\xi)\approx 0$,却不是最小点)且平台"面积"较大时,优化搜索过程一旦陷入平台就徘徊不前了.采取 **RO 算法**可以逃逸出平台并找到全局最优点或质量较好的局部最优点.下面介绍的是各种 RO 算法中的一种.

步 1　设 ξ 的定义域为 D.任择一初值参数矢量 $\xi(0)\in D$.定义最大搜索次数为 K.令搜索节拍 $k=0$.

步 2　产生一个高斯随机矢量 $\gamma(k)$,其均值为 $b(k)$,其维数与 ξ 相同.当 $k=$

0时，$b(k)=0$.若$(\xi(k)+\gamma(k))\in D$，转入步3；否则，转入步4.

步3 若 $R_e(\xi(k)+\gamma(k))<R_e(\xi(k))$，则令

$$\xi(k+1)=\xi(k)+\gamma(k) \text{ 且令 } b(k+1)=0.4\gamma(k)+0.2b(k).$$

若 $R_e(\xi(k)+\gamma(k))>R_e(\xi(k))$，且 $R_e(\xi(k)-\gamma(k))<R_e(\xi(k))$，则令

$$\xi(k+1)=\xi(k)-\gamma(k) \text{ 且令 } b(k+1)=-0.4\gamma(k)+b(k).$$

否则，令

$$\xi(k+1)=\xi(k) \text{ 且 } b(k+1)=0.5b(k).$$

步4 若 $k=K$，结束.若 $k<K$，令 $k=k+1$，转入步2.

5. MBP与RO混合算法

首先设立一个小的阈值参数 $\varepsilon>0$.第一步用MBP算法进行学习.设$\Delta R_e(k)=R_e(\xi(k))-R_e(\xi(k+1))$是节拍 k 的经验风险下降量，对每一节拍 k，检验 $\Delta R_e(k)>\varepsilon$ 是否成立.若不成立，表明已进入一个局部最小点或平台.接着转入RO算法进行学习.在RO过程中检验经验风险函数的积累降低量是否超过 $G=\max\{\mu\cdot R_e(\xi(k)),\Delta R_e(\xi(k))\}$，其中 $R_e(\xi(k))$ 和 $\Delta R_e(\xi(k))$ 分别为前面的MBP算法终结时的经验风险函数值及其下降量，μ 可以在 $0.1\sim0.2$ 之间选择.如果超过了，表明已从一个局部最小点或平台转移到了另一个更优的局部最小点附近.然后再转入MBP学习.这一过程可重复进行，直至找到一个满意的局部最小点或全局最小点.

1.2.3 RBF网络的学习算法

RBF网络皆取两层结构，其输出层取线性函数((1-3)、(1-4)式)，其隐层一般取高斯型径向基函数((1-8)、(1-9)式或(1-10)、(1-11)式).以MISO 2层RBF为例，设输入为 N 维矢量 $\boldsymbol{X}$，隐层包含 N_1 个神经元，输出为 y.当RBF取(1-8)、(1-9)式形式时，$\boldsymbol{X}$ 至 y 的映射可表示为

$$y=\sum_{i=1}^{N_1}w_i^{(2)}(\sqrt{2\pi}\sigma_i^{(1)})^{-1}\exp\left(-\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{W}_i^{(1)}\|^2/(\sigma_i^{(1)})^2\right). \quad (1\text{-}41)$$

其中 $\boldsymbol{W}_i^{(1)}=(w_{i1}^{(1)},w_{i2}^{(1)},\cdots,w_{iN}^{(1)})$是隐层第 i 神经元的权矢量，$w_i^{(2)}$ 是隐层第 i 神经元至输出层的权值.

若训练集为$(\hat{y}_1,\boldsymbol{X}_1),\cdots,(\hat{y}_M,\boldsymbol{X}_M)$，则RBF网络的离线参数学习分下列三步进行：

步1 隐层各神经元参数 $\boldsymbol{W}_i^{(1)}$ 用LBG算法学习.LBG算法是用于矢量量化(VQ)的一种无监督聚类算法，亦称为**修正Loyld算法**或称为**K平均算法**.算法程序如下：

(1-1) 设置最大迭代次数 K 和终止门限 δ，设置 N 个初始权矢量 $\boldsymbol{W}_i^{(1)}(0)$，$i=1,2,\cdots,N$(在没有先验知识时，可设置其为零均值高斯随机矢量.更有效的设置方法见文献[2])，设 $k=-1$ 的编码误差 $D(-1)=\infty$，令迭代节拍 $k=0$.

(1-2) 以 $\boldsymbol{W}_i^{(1)}(k)$，$i=1,2,\cdots,N_1$，为中心将$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_M\}$划归 N_1 个聚类区 $\Omega_i(k)$，$i=1,2,\cdots,N_1$.划分准则是：

若 $\| X_m - W_i^{(1)}(k) \| \leqslant \| X_m - W_j^{(1)}(k) \|, \forall j \neq i$，则 $X_m \in \Omega_i(k)$.
这种划分方法称为最邻近划分.

(1-3)　以 $\Omega_i(k)$ 的"质心"作为 $W_i^{(1)}(k+1)$，即

$$W_i^{(1)}(k+1) = \frac{1}{M_i(k)} \sum_{X_m \in \Omega_i(k)} X_m. \tag{1-42}$$

其中 $M_i(k)$ 为 $\Omega_i(k)$ 中包含的 X_m 个数.

(1-4)　按下式计算编码误差 $D(k)$ 和相对编码误差下降值 $d(k)$：

$$D(k) = \sum_{i=1}^{N_1} \sum_{X_m \in \Omega_i(k)} \| X_m - W_i^{(1)}(k) \|,$$

$$d(k) = \frac{D(k-1) - D(k)}{D(k-1)}.$$

(1-5)　判断 $d(k) < \delta$?若回答为是，转入步(1-7)；否则转入步(1-6).

(1-6)　判断 $k < K$?若回答为是，令 $k = k+1$，转入步(1-2)；否则转入步(1-7).

(1-7)　结束.

上列程序中 δ 是一个远小于 1 的正数，(1-5) 步的判断成立时，表明继续进行叠代计算不会使编码误差有显著下降，故应结束计算. 如果结束计算时的 N_1 个聚类区及其质心分别用 Ω_i 和 $W_i^{(1)}$ 表示，则按照(1-8)、(1-9) 式计算的 RBF 中，$\sigma_i^{(1)}$ 可用下式计算：

$$(\sigma_i^{(1)})^2 = \frac{1}{M_i} \sum_{X_m \in \Omega_i} \| X_m - W_i^{(1)} \|^2, \quad i = 1,2,\cdots,N_1. \tag{1-43}$$

其中 M_i 是 Ω_i 中所含 X_m 个数. 按照(1-10)、(1-11) 式计算 RBF 时，其 $\Sigma_i^{(1)}$ 可用下式计算

$$\Sigma_i^{(1)} = \frac{1}{M_i} \sum_{X_m \in \Omega_i} (X_m - W_i^{(1)})^{\mathrm{T}} (X_m - W_i^{(1)}), \quad i = 1,2,\cdots,N_1. \tag{1-44}$$

步 2　在固定 $W_i^{(1)}$ 及 $\sigma_i^{(1)}$ 或 $\Sigma_i^{(1)}$ 的条件下，对于隐层至输出层的各个权值 $w_i^{(2)}, i = 1,2,\cdots,N_1$，用 BP 算法进行训练(有监督学习).

步 3　对整个网络用 BP 算法进行训练.

由于步 2 和步 3 的 BP 算法与前述 MLP 的 BP 算法相似，不赘述. 在完成这三步学习计算时均反复调用训练集的数据. RBF 网络的学习速度比 MLP 网络要快很多，因此更适用于在线(实时) 的情况.

1.3　MLFN 推广能力的统计学习理论

基于 ERM 原理求出的一个 MLFN 的最优参数 ξ_a，只能使针对于训练集 $(\widehat{y}_1, X_1), \cdots, (\widehat{y}_M, X_M)$ 的经验风险函数 $R_e(\xi)$ 达到最小. 问题是 ξ_a 能否使风险函数 $R(\xi)$ 也达到足够小. 如果 $R(\xi_a)$ 与 $R(\xi)$ 的下界 $R(\xi_0)$ 十分接近，这表明用有限的

训练集求得的 ξ_a 对训练集外的数据也适用,即网络有良好的推广能力.一个推广能力不佳的网络没有实用价值,因此推广问题是包括 MLFN 在内的所有学习机的一个最关键的课题.推广能力取决于 MLFN 的规模和训练集的大小,而且只能在 X 和 $\hat{y}$ 的概率分布函数均未知的条件下求推广能力界(bound).维普尼克(Vapnik)和切尔伏宁齐斯(Chervonenkis)系统研究和解决了推广能力界的一系列问题,提出了统计学习理论(见文献5).

1.3.1 学习机的推广能力界

定理1 设有一实现映射 $y = f(X,\xi), \xi \in \Lambda$ 的学习机,ξ 是一组可调参数.设待实现的映射是 $X \to \hat{y}$(用 $P(X)$ 和 $P(\hat{y}/X)$ 来表征,这两个概率分布函数是确定的但又是未知的).设损失函数 $L(\hat{y}, f(X,\xi))$ 的上、下界为 A、B,即

$$-\infty < B \leqslant L(\hat{y}, f(X,\xi)) \leqslant A < \infty.$$

设此学习机的最小风险为 $R(\xi_0)$(见(1-14)式和(1-15)式),而由 $(\hat{y}, X_1), \cdots, (\hat{y}_M, X_M)$ 训练集按 ERM 原理求出的最佳参数是 ξ_a(见(1-16)和(1-17)式).那么对于任何 $\eta, 0 < \eta \ll 1$, $R(\xi_a)$ 满足下列不等式的概率不小于 $(1-2\eta)$(参考文献5).

$$R(\xi_a) \leqslant R(\xi_0) + \frac{B-A}{2}\sqrt{\epsilon} + (B-A)\sqrt{\frac{-\ln\eta}{2M}}, \tag{1-45}$$

其中当 $M > h$ 时,

$$\epsilon = \frac{4\left[h\ln\left(\frac{2M}{h}+1\right) - \ln\frac{\eta}{4}\right]}{M}. \tag{1-46}$$

h 称为学习机的 VC dim(Vapnik and Chervonenkis dimension 的缩写),它取决于学习机的结构和规模.对于一个隐层取 Sigmoid 函数、输出层取线性函数的2层 MISO-MLFN,若输入矢量 X 的维数为 N,隐层含 N_1 个神经元,则其 h 满足下列不等式

$$4\left[\frac{N_1}{2}\right]_I \cdot N \leqslant h \leqslant 4N_w \lg(eN_N), \tag{1-47}$$

其中 $[\cdot]_I$ 表示取整数部分,N_w 是网络中权的个数,N_N 是网络中神经元的个数(不包含输入层);$N_w = [(N+2)N_1 + 1]$, $N_N = N_1 + 1$.若 $N \gg 1$ 且 $N_1 \gg 1$,则上列不等式可以近似表示为

$$2N_1 N \leqslant h \leqslant 4N_1 N \lg(eN_1).$$

一个在实际应用中有效的经验公式是

$$h \approx 4N_1 N. \tag{1-48}$$

由(1-45)式可知,为使 $R(\xi_a)$ 足够小,必须使该式右侧的三项都足够小.下面先讨论如何计算 MLFN 的 $R(\xi_0)$.

1.3.2 MLFN 的函数逼近能力及 $R(\xi_0)$ 的计算

设一个有待于 MLFN 实现的映射是 $y = \hat{f}(X), X \in \mathscr{X} \subset \mathbf{R}^N$,而该 MLFN 所实

现的函数 $y = f(X)$ 可以在下列两种意义上逼近 $y = \hat{f}(X)$.

(1) **均匀逼近**　设

$$\|f - \hat{f}\|_u = \sup_{X \in \mathscr{D}} |f(X) - \hat{f}(X)|.$$

若对于任何 $\epsilon > 0$，总能设计 MLFN 的隐层数和各隐层神经元数，使得 $\|f - \hat{f}\|_u < \epsilon$ 成立，则称其具有均匀逼近能力.

(2) **均方逼近**　设

$$\|f - \hat{f}\|_{L^2} = \int_{\mathscr{D}} |f(X) - \hat{f}(X)|^2 \mathrm{d}X.$$

若对于任何 $\epsilon > 0$，总能设计 MLFN 的隐层数及各隐层神经元数，使得 $\|f - \hat{f}\|_{L^2} < \epsilon$ 成立，则称其具有均方逼近能力.

1. 2 层单输出 MLP 的函数逼近能力

对于隐层取 Sigmoid 函数、输出层取线性函数的一个 2 层 MISO-MLP，若隐层含 N_1 个神经元，则其所实现的映射 $y = f(X)$ 可表示为

$$y = \sum_{i=1}^{N_1} w_i^{(2)} \quad f_s((W_i^{(1)} \cdot X + \theta_i^{(1)})) + \theta^{(2)}, \tag{1-49}$$

其中 $f_s(\cdot)$ 为(1-6)式的 Sigmoid 函数，也可以用(1-7)式的形式. 此网络的函数逼近能力由下列定理给出.

定理 2(Cybenko 定理)　对于任何有限支单值连续函数 $y = \hat{f}(X)$，只要 N_1 足够大，就能用此网络实现均匀逼近或均方逼近.

2. RBF 网络的函数逼近能力

RBF 网络所实现的映射 $y = f(X)$ 由(1-41)式给出，它的函数逼近能力与上述 Cybenko 定理给出的结果一致.

3. $R(\xi_0)$ 的计算

$R(\xi_0)$ 取决于被逼近函数 $y = \hat{f}(X)$ 的特性和 MLFN 的结构及规模. 为描述被逼近函数的特性，方法之一是利用 $\hat{f}(X)$ 的傅里叶变换 $\hat{F}(\omega)$，两者具有如下关系

$$\hat{f}(X) = \int_{\mathbf{R}^N} e^{jX \cdot \omega} \hat{F}(\omega) \mathrm{d}\omega, \tag{1-50}$$

其中 $\omega = (\omega_1, \omega_2, \cdots, \omega_N) \in \mathbf{R}^N$. 定义参数 $C_{\hat{f}}$ 如下

$$C_{\hat{f}} = \int_{\mathbf{R}^N} \|\omega\| \cdot \|\hat{F}(\omega)\| \mathrm{d}\omega. \tag{1-51}$$

注意，$j\omega \hat{F}(\omega)$ 是 $\nabla_X \hat{f}(X)$ 的傅里叶变换. $C_{\hat{f}}$ 表征 $\hat{f}(X)$ 随 X 而变化的剧烈程度.

定理 3(Barron 定理)　设有一个 2 层 MISO-MLP，其映射特性用(1-49)式表示. 设 $\hat{f}(X)$ 满足下列两点要求：第一，X 的定义域 $\mathscr{D}$ 包含 $X = 0$ 且处在一个半径为 r 的超球 B_r(球心为 0)之内. 第二，参数 $C_{\hat{f}}$ 为有限值. 对于任何 $N_1 \geq 1$，总可以找到一组最佳权参数 ξ_0 使下式成立：

$$\int_{B_r}[\hat{f}(X)-f(X,\xi_0)]^2\mathrm{d}P(X)\leqslant\frac{(2rC_{\hat{f}})^2}{N_1},\tag{1-52}$$

其中 $f(X,\xi)$ 表(1-49)式右侧，ξ_0 是使(1-52)式左侧达到最小的权参数(包括阈值参数)，$P(X)$ 是 X 的概率分布函数. 此外，下列条件成立

$$\theta^{(2)}=\hat{f}(0),\quad \sum_{i=1}^{N_1}|w_i^{(2)}|\leqslant 2rC_{\hat{f}}.\tag{1-53}$$

注意，(1-52)式左侧即是按照回归估计损失函数(即误差平方，见(1-13)式)定义的最小风险函数，因此该式可表示为

$$R(\xi_0)\leqslant\frac{(2rC_{\hat{f}})^2}{N_1}.\tag{1-54}$$

当 $N_1\to\infty$ 时，$R(\xi_0)\to 0$ 对任何 $P(X)$ 成立.

1.3.3 结构风险最小原理

由(1-45)式可知，如果训练集的规模 M 不受限制，那么只要选足够大的 M 就可以使该式右侧第二、三两项充分小；同时只要 N_1 选得足够大，就可以使该式右侧第一项充分小. 这样，对于 M 和 N_1 不受限制的学习过程，$R(\xi_a)$ 可以任意小. 但是在实际中 M 不可能无限增大，这时学习过程必须将 M 固定在一可能的最大值上，而 N_1 成为唯一可供选择的参数. 如果 M 为某个固定值，则根据(1-46)和(1-48)式可得(1-45)式右侧第二、三项之和将只是 N_1 的函数(设 N 和 η 为固定值)，可以将其表示为 $T(N_1)$. 再利用(1-54)式，则(1-45)式可以重写为下列形式：

$$R(\xi_a)\leqslant\frac{(2rC_{\hat{f}})^2}{N_1}+T(N_1).\tag{1-55}$$

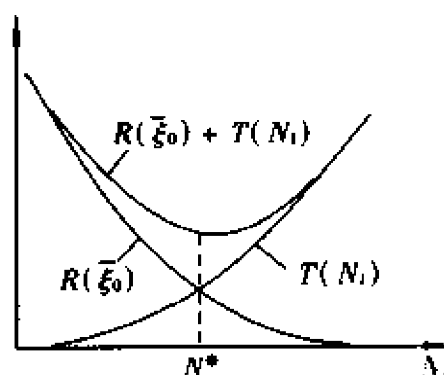


图 1-2

$T(N_1)$ 称为置信限，它随 N_1 的增加而增加. 而(1-55)式右侧第一项随 N_1 的增加而减少. 因此必然存在一最佳值 N_1^* 使这两项之和为最小，如图1-2所示. 按照这一原则来选择 N_1 值的学习机称为采用了结构风险最小原理.

1.4 改善 MLFN 推广能力的实用方法

虽然按照统计学习理论和 MLFN 函数逼近能力的定理，在 M 一定的条件下可以求出2层 MISO-MLP 的最佳隐层神经元个数 N_1^*，但是在实际中求 N_1^* 是困难的. 因为待逼近的函数 $\hat{f}(X)$ 是未知的，因而 $C_{\hat{f}}$ 难以求得. 此外，理论上也不能确定，就一个具体问题而言，二层、三层甚至更多层 MLFN 中何者为佳. 因此需要易于实现的在实际中改善 MLFN 推广能力的算法. 本节介绍其中的一部分.

1.4.1　交叉有效法

将所有采集到的训练数据$(\hat{y}_m, X_m)$分成两部分，一部分称为训练集，另一部分称为测试集. 在用BP算法按节拍k对网络进行训练时，只用训练集中的数据，每一节拍k的训练集内经验风险用$R_{T_{ee}}(k)$表示，而用测试集中的数据求得的训练集外经验风险用$R_{T_{te}}(k)$表示. 在训练时，$R_{T_{te}}(k)$和$R_{T_{ee}}(k)$随k的变化如图1-3所示. 在开始阶段二者皆随k的增加而下降，而当$k > k^*$后，$R_{T_{ee}}(k)$仍然下降，$R_{T_{te}}(k)$却上升了. 这说明，虽然继续训练会使集内经验风险进一步减小而推广效果反而变坏，这称为过适应. 因此，学习过程应在$k = k^*$时停止.

1.4.2　规格化方法

按照解决病态问题的理论和技术，用下列函数$\tilde{R}_e(\xi)$代替经验风险函数$R_e(\xi)$来求最佳参数ξ_a，可以有效地提高网络的推广性能.

$$\tilde{R}_e(\xi) = R_e(\xi) + \gamma\Omega(\xi), \tag{1-56}$$

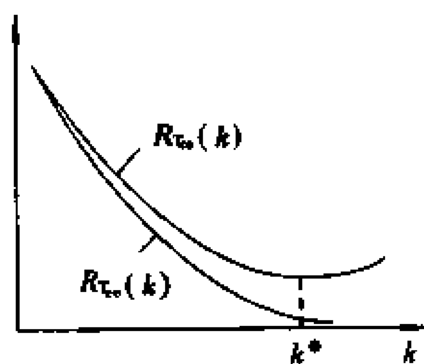


图 1-3

其中$\gamma\Omega(\xi)$称为**规格化项**. 按照对其不同选择可以形成不同方案. 下面介绍其中几种.

1. 权衰减法

设规格化项

$$\gamma\Omega(\xi) = \gamma\sum_{i,j,l}(w_{ij}^{(l)})^2, \tag{1-57}$$

此式右侧包括网络中所有权的平方(阈值参数$\theta_i^{(l)}$也应包括，为简化而略去). 按照最陡下降算法，为求得使$\tilde{R}_e(\xi)$为最小的ξ_a，按节拍k进行递推计算时权调整量$\Delta w_{ij}^{(l)}(k)$应由(1-20)式改为

$$\Delta w_{ij}^{(l)}(k) = \left[-\alpha\frac{\partial R_e(\xi)}{\partial w_{ij}^{(l)}} - 2\alpha\gamma w_{ij}^{(l)}\right]\Bigg|_{\xi=\xi(k)}, \tag{1-58}$$

其中$\gamma > 0$，称为规格化系数，$\alpha > 0$为已知步幅. 这种方案的特点是每迭代一步，各个权值的绝对值就会减小一点，衰减量的大小取决于γ的选值和权的幅度(绝对值)，幅度越大则衰减量越大，或称为受惩罚越严重. 这种方案使网络中的各权值幅度趋向于较小的值. 当某个权值幅度低于一个预定的小门限值时，即可将其删除，从而达到缩小网络规模的目的. γ的大小应仔细选择，γ过大将使各权值偏小，反之则不起作用. 权衰减法尚可作下列两种修正：第一种，将规格化项改为

$$\gamma\Omega(\xi) = \gamma\sum_{i,j,l}|w_{ij}^{(l)}|, \tag{1-59}$$

这时(1-58)式右侧第二项将改变为$-2\alpha\gamma\,\mathrm{sign}|w_{ij}^{(l)}|$；其优点是大幅度与小幅度权值受到相同的惩罚. 第二种，将规格化项改为$\sum_{i,j,l}\gamma_{ij}^{(l)}(w_{ij}^{(l)})^2$，其中$\gamma_{ij}^{(l)} =$

$\gamma[(w_{ij}^{(l)})^2]^{-p}$. $p=0$,相当于(1-57);$p=\frac{1}{2}$,相当于(1-59)式;$p=\frac{1}{3}$,则为二者之间的情况;$p=\frac{2}{3}$,则大幅度的权值受到的惩罚反而小于小幅度权值.

2. 只惩罚小幅度权值的方案

规格化项为

$$\gamma\Omega(\xi)=\gamma\sum_{i,j,l}\frac{(w_{ij}^{(l)}/w_0)^2}{1+(w_{ij}^{(l)}/w_0)^2}. \tag{1-60}$$

当$(w_{ij}^{(l)}/w_0)^2\gg 1$时,上列求和式中的对应项接近于 1,则该项不受惩罚;当$(w_{ij}^{(l)}/w_0)^2\ll 1$时,则对应项接近于$(w_{ij}^{(l)}/w_0)^2$,与前述(1-57)式一致.这样,只有幅值较 w_0 小得多的权受到惩罚,反之则惩罚较轻.

1.4.3 删除法

改善网络推广性能的最重要途径是采用规模尽可能小的网络.可以先取一个规模较大的网络,然后将那些"不重要"的权和神经元逐渐删除掉,最终达到缩小网络规模的目标.所谓不重要是指某个权或神经元被删除后,经验风险函数 $R_e(\xi)$ 没有明显变化.这就是说,$R_e(\xi)$ 相对于一个权的敏感度越低,则该权越应被删除.所以这种方法也称为敏感度法.按照对于敏感度的不同定义可以构成不同的删除方法,下面介绍其中的几种.

1. 关联度法

首先定义一个权的**关联度** ρ,ρ 为该权删除后与删除前的$R_e(\xi)$之差值.ρ 即可作为该权敏感度的度量.由于直接计算 ρ 时开销太大,可以采用一种间接的方法.首先,将每个神经元的函数 $x_i^{(l)}=f_l\left(\sum_{j=1}^{N_{l-1}}w_{ij}^{(l)}x_j^{(l-1)}+\theta_i^{(l)}\right)$ 改变为 $x_i^{(l)}=f_l\left(\sum_{j=1}^{N_{l-1}}\zeta_{ij}^{(l)}w_{ij}^{(l)}x_j^{(l-1)}+\zeta_i^{(l)}\theta_i^{(l)}\right)$.这样,权 $w_{ij}^{(l)}$ 的关联度$\rho_{ij}^{(l)}$ 可以用下式近似计算:

$$\rho_{ij}^{(l)}\approx-\left.\frac{\partial R_e(\xi)}{\partial\zeta_{ij}^{(l)}}\right|_{\zeta_{ij}^{(l)}=1}. \tag{1-61}$$

事实上,此式可以在 BP 计算中予以实现.而且计算结果是在 $\zeta_{ij}^{(l)}=1$ 时估计的,所以在最终的计算公式中此参数并不出现.如果$|\rho_{ij}^{(l)}|$低于某一门限值,则可以将 $w_{ij}^{(l)}$ 删除.删除过程是首先进行正常的 BP 学习,在参数收敛到一个最佳值后删除若干个关联度低于门限的权值,然后再进行 BP 学习,这一过程可反复交替进行直至网络规模有了较大压缩为止.

此方法可以作两项修正:第一,在(1-61)式的计算中,$R_e(\xi)$(见(1-18)式)用下列 $R_A(\xi)$ 替代

$$R_A(\xi)=\frac{1}{M}\sum_{m=1}^{M}|f(X_m,\xi)-\hat{y}_m|, \tag{1-62}$$

而 BP 计算仍然采用 $R_e(\xi)$. 这可以使 $R_e(\xi)$ 值较小时关联度 $\rho_{ij}^{(l)}$ 的估计较直接用 $R_e(\xi)$ 更为准确. 第二, 为了避免随机性, 不是在 BP 学习结束时估计 $\rho_{ij}^{(l)}$, 而是在整个 BP 学习过程中用下列惯性算法进行估计:

$$\rho_{ij}^{(l)}(k+1) = 0.8\rho_{ij}^{(l)}(k) + 0.2\left(-\frac{\partial R_e(\xi)}{\partial \zeta_{ij}^{(l)}}\right)\Bigg|_{\substack{\zeta_{ij}^{(l)}=1 \\ \xi=\xi(k)}}. \tag{1-63}$$

2. 灵敏度法

设 BP 算法按节拍从 $k=0$ 开始到 $k=K$ 结束, 任意一个权 $w_{ij}^{(l)}$ 将从 $w_{ij}^{(l)}(0)$ 出发, 到 $w_{ij}^{(l)}(K)$ 终止. 这样, 可以用灵敏度 $S_{ij}^{(l)}$ 来表征整个 BP 学习过程中 $w_{ij}^{(l)}$ 的变化对于 $R_e(\xi)$ 的影响

$$S_{ij}^{(l)} = -\sum_{k=0}^{K-1}\frac{\partial R_e(\xi)}{\partial w_{ij}^{(l)}}\Bigg|_{\xi=\xi(k)} \cdot \Delta w_{ij}^{(l)}(k)\frac{w_{ij}^{(l)}(K)}{w_{ij}^{(l)}(K) - w_{ij}^{(l)}(0)}. \tag{1-64}$$

这样, 在 BP 学习结束时可以将灵敏度最低或低于某个门限的若干个权删掉, 然后再一次进行 BP 学习. 如此可反复进行若干次.

3. 突出特征法

假设网络已经通过了充分训练, 这时可以通过下列方法来确定 $R_e(\xi)$ 对某个权 $w_{ij}^{(l)}$ 的敏感程度. 设终止训练时网络的权参数为 $\xi(K)$, 这时 $w_{ij}{}^{(l)}$ 的变化量若为 $\Delta w_{ij}{}^{(l)}$, 则 $R_e(\xi)$ 的变化量 $\Delta R_e(\xi)$ 可用下式计算:

$$\Delta R_e(\xi) = \frac{\partial R_e(\xi)}{\partial w_{ij}{}^{(l)}}\Bigg|_{\xi=\xi(K)} \cdot \Delta w_{ij}{}^{(l)} + \frac{1}{2}\frac{\partial^2 R_e(\xi)}{\partial (w_{ij}{}^{(l)})^2}\Bigg|_{\xi=\xi(K)} \cdot [\Delta w_{ij}{}^{(l)}]^2 + \cdots.$$

由于网络已通过充分学习, 此式右侧第一项中的一阶偏导数十分接近于 0. 如果略去二阶交叉项和三阶及三阶以上偏导数的影响, 那么删除 $w_{ij}{}^{(l)}$ 所造成的 $R_e(\xi)$ 的变化主要由上式中的二阶偏导数项造成. 这样可以定义一个称为**突出特征**的参数 $T_{ij}{}^{(l)}$ 如下:

$$T_{ij}{}^{(l)} = \frac{\partial^2 R_e(\xi)}{\partial (w_{ij}{}^{(l)})^2}\Bigg|_{\xi=\xi(K)} \cdot [w_{ij}{}^{(l)}(K)]^2. \tag{1-65}$$

和前几种方法一样, 先对一个规模较大的网络进行训练, 然后删掉 $T_{ij}{}^{(l)}$ 最小的若干个权. 这一学习和删除交替进行.

$T_{ij}{}^{(l)}$ 涉及二阶偏导数的计算, 这也可以通过 BP 算法求得. 下面以(1-49)式表示的 2 层 MISO-MLP 为例说明其计算方法.

首先, 对于训练集中的每一组数据 $(\hat{y}_m, X_m)$, $m=1,2,\cdots,M$, 进行前向计算, 求得以下参数:

隐层各神经元输出:

$$x_{mi}{}^{(1)} = f_s(I_{mi}{}^{(1)}),\ I_{mi}{}^{(1)} = \sum_{j=1}^{N} w_{ij}{}^{(1)}x_{mj} + \theta_i{}^{(1)},\quad i=1,2,\cdots,N_1;$$

输出层神经元输出:

$$y_m = x_{m1}{}^{(2)} = \sum_{i=1}^{N_1} w_i{}^{(2)}x_{mi}{}^{(1)} + \theta^{(2)}.$$

引用(1-18)式，

$$R_e(\xi)=\frac{1}{M}\sum_{m=1}^{M}(\widehat{y}_m-y_m)^2=\frac{1}{M}\sum_{m=1}^{M}(\widehat{y}_m-x_{m1}^{(2)})^2.$$

这样可以求得

$$\frac{\partial^2 R_e(\xi)}{\partial(w_i^{(2)})^2}=\frac{2}{M}\sum_{m=1}^{M}(x_{mi}^{(1)})^2,\quad i=1,2,\cdots,N_1,$$

$$\frac{\partial^2 R_e(\xi)}{\partial(w_{ij}^{(1)})^2}=\frac{2}{M}\sum_{m=1}^{M}\Big([f'_S(I_{mi}^{(1)})]^2(w_i^{(2)})^2-(\widehat{y}_m-y_m)f''_S(I_{mi}^{(1)})w_i^{(2)}\Big)x_{mj}^2,\quad i=1,2,\cdots,N_1,\ j=1,2,\cdots,N,$$

$$f'_S(I_{mi}^{(1)})=x_{mi}^{(1)}(1-x_{mi}^{(1)}),$$

$$f''_S(I_{mi}^{(1)})=(1-2x_{mi}^{(1)})x_{mi}^{(1)}(1-x_{mi}^{(1)}).$$

1.4.4 交叉删除法

仍以(1-49)式表示的2层MISO-MLP为例来说明这种方法.对于其隐层各神经元的权矢量 $W_i^{(1)}$，如果某两个权矢量之间的"夹角"很小，表明它们的功能近似，因而其中的一个可以被看成多余的而删掉.具体作法是首先将各个 $W_i^{(1)}$ 写成下列形式：

$$W_i^{(1)}=\widehat{W}_i^{(1)}G_i^{(1)},$$

其中实数 $G_i^{(1)}>0$，$\|\widehat{W}_i^{(1)}\|=1$. $G_i^{(1)}$ 称为 $W_i^{(1)}$ 的增益.然后，在普通BP算法的每一步权参数调整之后再增加一步增益调整.增益调整量 $\Delta G_i^{(1)}$ 按下式计算：

$$\Delta G_i^{(1)}=-\mu\sum_{\substack{j=1\\ j\neq i}}^{N_1}(\widehat{W}_i^{(1)}\cdot\widehat{W}_j^{(1)})^2G_j^{(1)},\ i=1,2,\cdots,N_1,\qquad(1\text{-}66)$$

其中$(\widehat{W}_i^{(1)}\cdot\widehat{W}_j^{(1)})$即等于 $W_i^{(1)}$ 和 $W_j^{(1)}$ 之间夹角的余弦，当两个矢量相似度越高时，此值越接于1.当某个隐层神经元 i 的增益 $G_i^{(1)}$ 调到接近于0或小于0时，该神经元即被删除.在实际执行此算法时可以进行若干步BP权调整算法后，再进行一次增益调整.权调整步幅 $\mu>0$，其值必须选择恰当. μ 太小起不到删除作用，μ 太大则删除过多.

1.4.5 局部连接和权分享

在图像识别、文字和符号识别以及语音识别等领域中，输入的空域或时域信号的某个特征往往只出现在局部区域中，而该局部区域又可能处于全域的任何一个位置.这样，MLFN第一隐层的各个神经元不必与表示全局特征的输入矢量 X 的每个分量相连接，而只需与局部若干分量相连接.例如，一个分辨率为(16×16)的图像中包含256个像素，它可以构成输入一个MLFN的256维矢量 X.若某些局部特征只反映在一个(5×5)的局部区域中，这时第一隐层的一个神经元只需与 X 中的25

个分量相连接,从而大大减少了权的数量.从第一隐层向更高隐层传送信号时也可照此办理.图1-4给出了一个2层MISO-MLP的示例,其中输入是一个4维矢量 $X=(x_1,x_2,x_3,x_4)$(输入层还有一个为提供阈值参数的输出为1的神经元),隐层有6个神经元(另有一个输出为1的神经元以提供阈值参数),输出层只有一个神经元.隐层神经元分成(A,B,C)和(D,E,F)两组.A只与x_1、x_2相连,B与x_2、x_3相连,C与x_3、x_4相连;D、E、F的连接与此同.A、B、C取相同输入权值,它们用来检测同一种局部特征;D、E、F取另一组相同输入权值,用来检测另一种局部特征.这就是局部连接和权分享名称的由来.如果此网络取全连接结构,将使用37个权参数(包括阈值参数);而使用此结构时将包含25个权(包括阈值参数),由于权分享,可自由调节的权参数是13个.在实际中网络的规模一般很大,采用局部连接和权分享可以大大减少自由权参数的数量,从而有效地改善其推广性能.

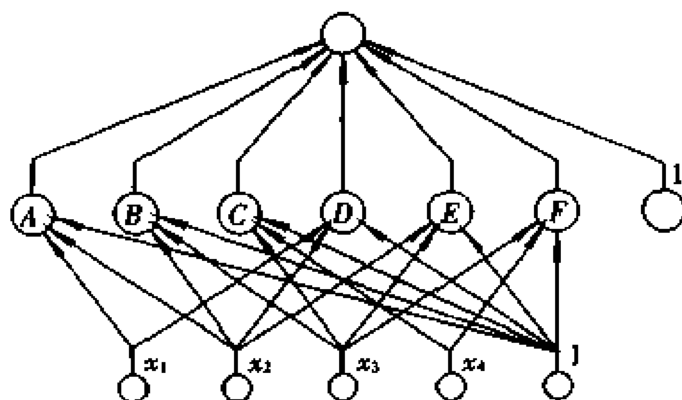


图 1-4

1.4.6　局部优化法

一个待逼近的函数 $y=\hat{f}(X)$ 在 X 的定义域 $\mathscr{D}$ 的全局中的变化既可能很大又很复杂,而在一个较小的局部区域中就可能简单得多.图1-5中给出了 X 为一维时,$y=\hat{f}(x)$ 的一个示例,其中 $\mathscr{D}=[x_1,x_5]$.若将 x 分为4个小区$[x_1,x_2)$、$[x_2,x_3)$、$[x_3、x_4)$、$[x_4、x_5]$,则每个小区中 $\hat{f}(x)$ 的变化较简单.

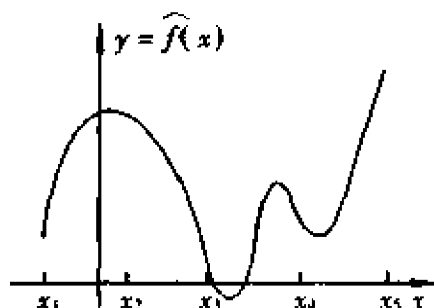


图 1-5

这样，可以将 X 的定义域 $\mathscr{D}$ 分为若干个小区，每个小区单独用一个较简单的 MLFN 来实现所需的映射. 区域划得越细，所需的各小区 MLFN 越简单，因而有可能改善推广性能. 但是当训练集规模 M 一定时，小区划分越细时每个小区分得的训练数据越少，这又可能恶化推广性能. 此外，过多设置小区也会使计算开销过大. 所以只能作折衷的选择.

划分小区除了用上面提到的等间隔划分法外. 另一种效果更好且更有实用价值的方法是采用，1.2.3节所述的LBG算法，将 $\mathscr{D}$ 分为若干个聚类区，每区有一个中心；当 X 与某区中心最接近时就划归该区并用该区的 MLFN 实现所需的映射.

1.4.7　多网络法

1. 模式识别情况

可以用 $(2Q+1)$ 个 MLFN 来代替单个 MLFN 实现模式识别任务，其中 Q 为某个正整数. 这 $(2Q+1)$ 个网络可以用下列各种方案分别进行学习：

(1) 各网络结构相同且用同样的训练集，但是采用不同的权参数初值进行训练，学习结束时各网络具有不同的权参数，它们相应于 $R_r(\xi)$ 的不同局部最小点.

(2) 将训练集分成 $(2Q+1)$ 个子集，每个网络用不同的子集进行训练，从而导致不同的权参数.

(3) 每个网络的训练集和参数初值相同，但是其结构有差异(例如隐层神经元数不同)，从而导致不同的网络.

(4) 以上几种方案的组合. 在识别时(以二类划分为例) $(2Q+1)$ 个网络可以产生 $(2Q+1)$ 个结果，每个结果不是1就是0，这时可以用少数服从多数的表决法给出最终识别结果. 实验结果表明，这种方法的确可以改善网络的推广性能，即明显地降低集外误识率. 其代价是增大学习开销，并且使识别的计算量和存储量增加.

2. 函数逼近情况

这时可以用 Q 个 MLFN 实现所需的映射 $y=\hat{f}(X)$，Q 为任何正整数. 每个网络的结构和训练仍可按照 1. 中所述的方法实施. 在应用时将各个网络的输出取算术平均作为所需输出. 这使得均方意义的推广效果优于单个网络.

3. 用多个 MISO 网络替代单个 MIMO 网络

如待实现的映射是 $X \rightarrow Y$，其中 Y 是一个维数大于1的矢量(在模式识别和函数逼近领域都有这种情况)，一般可以用单个 MIMO 网络来实现. 如果 Y 的维数为 N_L，也可以用 N_L 个 MISO 网络来实现，每个网络的输出等于 Y 的一个分量. 这种方案也可以减小每个网络的容量从而改善其推广性能.

1.5　MLFN 作为后验概率估值器

设输入矢量 X 与 L 个类别 $C^{(l)}$，$l=1,2,\cdots,L$，相联系，它们之间的关系用联合概率分布函数 $P(X,C^{(l)})$ 描述. $P(X,C^{(l)})=P(X)P(C^{(l)}/X)$.

正确估计后验概率 $P(C^{(l)}/X)$ 是一个既有用而又困难的任务. 现在基于一个

训练集(C_m, X_m), $m = 1,2,\cdots,M$, 来对一个 MLFN 进行训练，其中 $C_m \in \{C^{(1)}, C^{(2)}, \cdots, C^{(L)}\}$；MLFN的输入为 X，输出为 $Y = (y_1, y_2, \cdots, y_L)$. 在训练时，若输入为 X_m，则理想输出 $\hat{Y}_m = (\hat{y}_{m1}, \hat{y}_{m2}, \cdots, \hat{y}_{mL})$；当 $C_m = C^{(l)}$，则 $\hat{y}_{mi} = \delta_{il}$，即当 $i = l$ 时，有 $\delta_{il} = 1$，当 $i \neq l$ 时，有 $\delta_{il} = 0$. 完成训练后此网络可以用作分类器，即每输入一个 X 时以网络的 L 个输出中的最大输出端标号作为 X 所属的类别编号. 下面将证明，在一定的条件下这 L 个输出值即等于 L 个后验概率 $P(C^{(l)}/X)$, $l = 1,2,\cdots,L$.

若 MLFN 的 L 个输出用 $y_i(X,\xi)$, $i = 1,2,\cdots,L$, 表示，则其经验风险函数 $R_e(\xi)$ 可以取两种形式. 第一种是均方误差形式：

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{L}[\hat{y}_{mi} - y_i(X_m,\xi)]^2; \tag{1-67}$$

第二种是相对熵函数形式：

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{L}\hat{y}_{mi}\ln\frac{\hat{y}_{mi}}{y_i(\hat{X}_m,\xi)}. \tag{1-68}$$

无论对于哪一种形式都可以证明，只要网络的权足够多以及所求得的 ξ_A 能使 $R_e(\xi)$ 达到全局最小，那么以 ξ_A 为参数的网络可以实现上述的后验概率估计. 现在仅就第一种情况予以证明. 首先，对任何 X_m 下式成立：

$$\sum_{l=1}^{L}P(C^{(l)}/X_m) = 1. \tag{1-69}$$

这样，(1-67) 式可以重写如下：

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{i=1}^{L}[\hat{y}_{mi} - y_i(X_m,\xi)]^2 P(C^{(l)}/X_m).$$

注意当 $C_m = C^{(l)}$ 时 $\hat{y}_{mi} = \delta_{il}$，上式可以表示为

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}\Big\{\sum_{i=1}^{L}[1 - y_i(X_m,\xi)]^2 P(C^{(i)}/X_m) + \sum_{i=1}^{L}[0 - y_i(X_m,\xi)]^2\Big[\sum_{\substack{l=1\\ l\neq i}}^{L}P(C^{(l)}/X_m)\Big]\Big\}.$$

由(1-69) 式可得

$$\sum_{\substack{l=1\\ l\neq i}}^{L}P(C^{(l)}/X_m) = 1 - P(C^{(i)}/X_m).$$

将此式代入前式并作简单推导，即得

$$R_e(\xi) = \frac{1}{M}\sum_{m=1}^{M}\Big\{\sum_{i=1}^{L}[y_i^2(X_m,\xi) - 2y_i(X_m,\xi)P(C^{(i)}/X_m) + P^2(C^{(i)}/X_m)]\Big\} + \frac{1}{M}\sum_{m=1}^{M}\Big\{\sum_{i=1}^{L}[P(C^{(i)}/X_m) - P^2(C^{(i)}/X_m)]\Big\}.$$

此式右侧第二个和式与 ξ 无关，因此只有当某个 ξ_A 使该式右侧第一个和式达到最

小时，$R_e(\xi)$ 才能达到最小．易于看到，只有当下列条件成立时才能实现这一要求：

$$y_i(X_m,\xi_A) = P(C^{(i)}/X_m), \quad m = 1,2,\cdots,M. \tag{1-70}$$

显然，当网络具有足够规模从而能提供所需的 ξ_A 且 $R_e(\xi_A)$ 为全局最小时，网络的第 i 端输出即等于后验概率 $P(C^{(i)}/X_m)$（网络输入为 X_m）．如果训练集的规模足够大，这一结果可以推广到 X 的定义域 $\mathscr{D}$ 中的所有矢量．

1.6　递归神经网络(RNN)

前述各节讨论的都是单个输入 X 对单个输出 Y 的静态映射．在非线性动力系统的辨识、控制、故障诊断以及时间序列预测等许多领域中，都涉及两个离散时间序列 $X(m)$ 和 $Y(m)$ 之间的动态映射，即某个离散时刻 k 的输出矢量 $Y(m)$ 不仅依赖于 $X(m)$，而且依赖于 $X(m-1)$，$X(m-2)$，… 以及 $Y(m-1)$，$Y(m-2)$、…．为了实现这种依赖关系，应该用图 1-6 所示的结构．

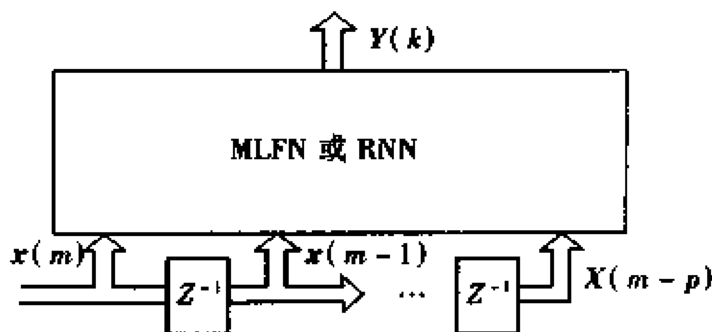


图 1-6

当其中的神经网络为 MLFN 时，$Y(m)$ 将依赖于 $X(m)$，$X(m-1)$，…，$X(m-p)$，它与线性系统理论中的 MA 模型对应，称为 NMA 模型（N 表示非线性）．当其中的神经网络为 RNN 时，$Y(m)$ 不仅依赖于上列各 X，而且依赖于 $Y(m-1)$、$Y(m-2)$，…，它与线性系统理论中的 ARMA 模型对应，称为 NARMA 模型．RNN 是在 MLFN 的基础上构成的，除了 MLFN 中的由入向出的单向信息传送通道外，每个隐层还包括自身的反馈通道．

在 NMA 模型中由于不存在反馈，其学习算法与一般 MLFN 无差异．而 NARMA 模型的学习算法必须在一般 MLFN 的 BP 学习算法框架下进行修正．下面以一个 L 层全连接 RNN 为例予以说明．此网络输出层（$l = L$）有 N_L 个神经元，皆取线性函数；$L-1$ 个隐层（$l = 1,2,\cdots,L-1$）各有 N_l 个神经元，皆取 Sigmoid 函数；输入层（$l = 0$）有 N_0 个神经元，其取值 $x_i^{(0)}$，$i = 1,2,\cdots,N_0$，与输入信号对应．网络中除有由入至出的信号传送途径外，每个隐层的各神经元之间存在全反馈连接，输出层不存在反馈，如图 1-7 所示．假设网络按离散时间 m 运行，各输出皆延时 1 至 $q^{(l)}$ 个节拍（图中用 Z^{-1} 表示一个节拍的延时）再反馈到同层的所有神经元．下面只研究每个隐层仅有一个节拍延时反馈的情况，即 $q^{(l)} = 1$，$l = 1,2,\cdots,L-1$．注意，网络的训

练集是两个对应的时间序列：$X^{(0)}(m),\hat{X}^{(L)}(m),m=1,2,\cdots$.

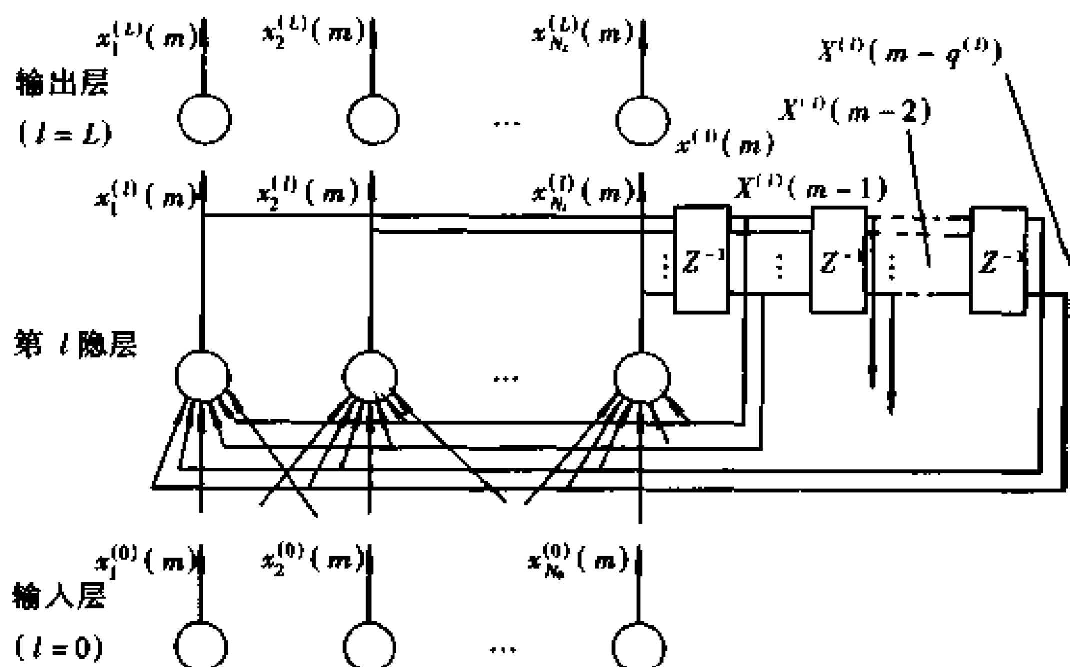


图 1-7

网络的运行过程如下.按照离散时间 $m=1,2,\cdots$ 向网络输送矢量 $X^{(0)}(m)$，且假设当 $m\leqslant 0$时网络中所有神经元输出 $x_i^{(l)}(m)=0$.这样，对于 $m\geqslant 1,x_i^{(l)}(m)$ 可用下式计算(注意，$w_{ij}^{(L,L)}\equiv 0,\forall\, i,j$)：

$$x_i^{(l)}(m)=f_l(I_i^{(l)}(m)),l=1,2,\cdots,L,\quad i=1,2,\cdots,N_l,m=1,2,\cdots,$$

$$I_i^{(l)}(m)=\sum_{j=1}^{N_l}w_{ij}^{(l,l)}\cdot x_j^{(l)}(m-1)+\sum_{j=1}^{N_{l-1}}w_{ij}^{(l,l-1)}\cdot x_j^{(l-1)}(m)+\theta_i^{(l)},\tag{1-71}$$

其中 $f_l(\cdot),l=1,2,\cdots,L-1$,为 Sigmoid 函数，$f_L(\cdot)$ 为线性函数.此式右侧第一个和式是第 l 层各神经元输出对本层第 i 神经元延迟一拍的反馈输入，第二个和式是第 $l-1$ 层各神经元对 l 层 i 神经元的前向输入，$\theta_i^{(l)}$ 是阈值参数.

网络的学习过程如下：给定训练集$\{X^{(0)}(m),\hat{X}^{(L)}(m)\mid m=1,2,\cdots\}$，其中 $\hat{X}^{(L)}(m)=(\hat{x}_1^{(L)}(m),\hat{x}_2^{(L)}(m),\cdots,\hat{x}_{N_L}^{(L)}(m))$ 是 m 时刻的网络理想输出，$X^{(0)}(m)=(x_1^{(0)}(m),x_2^{(0)}(m),\cdots,x_{N_0}^{(0)}(m))$是 m 时刻的网络输入.m 时刻的网络实际输出记为 $X^{(L)}(m)=(x_1^{(L)}(m),x_2^{(L)}(m),\cdots,x_{N_L}^{(L)}(m))$，它是 $X^{(0)}(m)$ 以及网络参数 ξ 的函数，为简洁起见，没有把这些符号标记出来.网络在 m 时刻的瞬时风险函数用 $R_e^{(m)}(\xi)$ 表示，其计算公式如下列：

$$R_e^{(m)}(\xi) = \sum_{i=1}^{N_L}[\hat{x}_i^{(L)}(m) - x_i^{(L)}(m)]^2. \tag{1-72}$$

网络的学习可以按照批处理(离线)和实时处理(在线)两种方式进行.按批处理方式时,首先采集 M 组训练数据 $\{\hat{X}^{(L)}(m), X^{(L)}(m) \mid m = 1,2,\cdots,M\}$ 构成的训练集.风险函数 $R_e(\xi)$ 定义如下:

$$R_e(\xi) = \sum_{m=1}^{M} R_e^{(m)}(\xi). \tag{1-73}$$

然后求最佳 ξ_A 使 $R_e(\xi)$ 达到最小.按实时处理方式时,边采集数边进行训练,这时风险函数表示为如下随 m 而改变的函数:

$$R_e(\xi, m) = \sum_{m'=m-\eta}^{m-1} R_e^{(m')}(\xi) \quad (\eta > 0).$$

此时,针对不同的 m 值求出相应的最佳参数 $\xi_A(m)$,使 $R_e(\xi, m)$ 达到最小.这样的最佳参数随时间而改变,从而能对变化的环境进行自适应.参数 $\eta > 0$ 越大,则求和范围越宽,因而使统计平均效果越好,η 越小,则网络的自适应效果越好.对于一个具体问题应作折衷选择.这两种方式的学习方法相似,下面以批处理方式为例来阐明.

在 BP 算法的框架下,按最陡下降算法,从随机设置的初始权值 $w_{ij}^{(l,l-1)}(0)$ 和 $w_{ij}^{(l,l)}(0)$ 出发,以节拍 k 进行迭代计算 $w_{ij}^{(l,l-1)}(k+1) = w_{ij}^{(l,l-1)}(k) + \Delta w_{ij}^{(l,l-1)}(k)$ 以及 $w_{ij}^{(l,l)}(k+1) = w_{ij}^{(l,l)}(k) + \Delta w_{ij}^{(l,l)}(k), k = 0,1,\cdots$,其中 $\Delta w_{ij}^{(l,l-1)}(k)$ 为前向权的调节量,其计算与前述静态网络情况相同(见(1-20)～(1-32)式).反馈权的调节量 $\Delta w_{ij}^{(l,l)}(k)$ 根据(1-71)式和(1-72)式计算如下:

$$\Delta w_{ij}^{(l,l)}(k) = -\alpha \sum_{m=1}^{M} \frac{\partial R_e^{(m)}(\xi)}{\partial w_{ij}^{(l,l)}} \Big|_{\xi = \xi(k)},$$

$$\frac{\partial R_e^{(m)}(\xi)}{\partial w_{ij}^{(l,l)}} = -2\sum_{q=1}^{N_L}[\hat{x}_q^{(L)}(m) - x_q^{(L)}(m)] \frac{\partial x_q^{(L)}(m)}{\partial w_{ij}^{(l,l)}}, \tag{1-74}$$

其中 α 为步幅,$0 < \alpha \ll 1$.这样,只要能对于任何 l'、l、q、m、i、j 计算 $\frac{\partial x_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}}$,就能将 $\Delta w_{ij}^{(l,l)}(k)$ 计算出来.引用(1-71)式,可得

$$\frac{\partial x_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}} = \frac{\mathrm{d}x_q^{(l')}(m)}{\mathrm{d}I_q^{(l')}(m)} \cdot \frac{\partial I_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}}.$$

此式右侧第一项的计算:当 $l' = L$ 时,$f_{l'}(\cdot)$ 为线性函数,它等于1;当 $l' = 1 \sim L-1$ 时,$f_{l'}(\cdot)$ 为 Sigmoid 函数,它等于 $x_q^{(l')}(m)(1 - x_q^{(l')}(m))$(见1.2.1节(1-27)式).右侧第二项可分三种情况计算如下.

(1) 若 $l > l'$,

$$\frac{\partial I_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}} = 0. \tag{1-75}$$

(2) 若 $l = l'$,

$$\frac{\partial I_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}} = \sum_{p=1}^{N_l}\left[w_{qp}^{(l,l)} \frac{\partial x_p^{(l)}(m-1)}{\partial w_{ij}^{(l,l)}} + \delta_{qi} x_j^{(l)}(m-1)\right], \tag{1-76}$$

其中 $\delta_{qi} = 1$,当 $q = i$;$\delta_{qi} = 0$,当 $q \neq i$.

(3) 若 $l < l'$,

$$\frac{\partial I_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}} = \sum_{p=1}^{N_{l'}}\left[w_{qp}^{(l',l')} \frac{\partial x_p^{(l')}(m-1)}{\partial w_{ij}^{(l,l)}}\right] + \sum_{r=1}^{N_{l'-1}}\left[w_{qr}^{(l',l'-1)} \frac{\partial x_r^{(l'-1)}(m)}{\partial w_{ij}^{(l,l)}}\right]. \tag{1-77}$$

这样只要对于所有 l'、l、i、j 给定初始条件 $\frac{\partial x_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}} = 0, m \leqslant 0$,就可以按时序 $m = 1,2,\cdots$ 依次算出各 $\frac{\partial x_q^{(l')}(m)}{\partial w_{ij}^{(l,l)}}$,再引用(1-74)式即可求得所需的反馈权调整量 $\Delta w_{ij}^{(l,l)}$. 在这一算法中,每一时序 m 的梯度(偏微分)计算都依赖于前一时序 $(m-1)$ 的梯度计算,故称为**动态梯度下降算法**. 若干实验结果表明,在非线性动力系统的辨识、控制以及时间序列预测等领域中,采用这种动态学习算法的 RNN 的性能优于静态的 MLFN.

2 Hopfield 神经网络

2.1 连续时间 Hopfield 神经网络

网络由 N 个神经元构成,按顺序编号为 $i = 1,2,\cdots,N$. 每个神经元 i 用 3 个变量来描述:输出 $x_i(t)$,状态 $u_i(t)$ 和输入 $I_i(t)$,其中 t 是连续时间变量. 对于任何 i 可以用下列方程表示其运算规则:

$$\begin{aligned} \frac{\mathrm{d}u_i(t)}{\mathrm{d}t} &= -\frac{u_i(t)}{\tau} + \sum_{j=1}^{N} w_{ij} x_j(t) + I_i(t), \\ x_i(t) &= f_i(u_i(t)), \quad i = 1,2,\cdots,N. \end{aligned} \tag{2-1}$$

其中 τ 是一正常数,w_{ij} 是第 j 神经元至第 i 神经元的信号传送权值且 $w_{ij} = w_{ji}$,所有权为实数. $f_i(\cdot)$ 是下列形式的单调非降函数:

$$f_i(u_i(t)) = \frac{1}{2}\left[1 + \tanh\left(\frac{u_i(t)}{u_0}\right)\right], \tag{2-2}$$

其中 u_0 是一正常数,当 $u_0 \to 0$ 时,$f_i(\cdot)$ 趋向于硬限幅函数(见(1-5)式). (2-1)式构成 N 元联立非线性方程组. 当给定初值 $x_i(0)$、$u_i(0)$ 和 $I_i(t)$, $t \geqslant 0$, $i = 1,2,\cdots,N$,就可以求出此方程组的解.

若 $I_i(t) = I_i(0), t \geqslant 0, i = 1,2,\cdots,N$. 对任何给定的 $I_i(0)$，当 $t \to \infty$ 时各 $x_i(t)$ 皆趋向某个固定值，则称系统具有稳定解. 下面用能量函数的方法确定稳定解存在的条件.

能量函数 $E(t)$ 定义如下：

$$E(t) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}x_i(t)x_j(t) + \frac{1}{\tau}\sum_{i=1}^{N}\int_0^{x_i(t)} f_i^{-1}(x)\mathrm{d}x - \sum_{i=1}^{N} I_i(0)x_i(t). \tag{2-3}$$

若 w_{ij}、$I_i(0)$、τ、u_0 为有限确定值($\tau > 0, u_0 > 0$)时，易于证明 $E(t)$ 为有下确界的函数. $E(t)$ 的导数可用下式求得：

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = \sum_{i=1}^{N}\frac{\partial E(t)}{\partial x_i(t)}\cdot\frac{\mathrm{d}x_i(t)}{\mathrm{d}t}. \tag{2-4}$$

由(2-3)式且注意到 $w_{ij} = w_{ji}$，可求得

$$\frac{\partial E(t)}{\partial x_i(t)} = -\sum_{j=1}^{N} w_{ij}x_j(t) + \frac{u_i(t)}{\tau} - I_i(0). \tag{2-5}$$

注意到 $I_i(t) = I_i(0)$，与(2-1)式对照，即得

$$\frac{\partial E(t)}{\partial x_i(t)} = -\frac{\mathrm{d}u_i(t)}{\mathrm{d}t} = -\frac{\mathrm{d}u_i(t)}{\mathrm{d}x_i(t)}\cdot\frac{\mathrm{d}x_i(t)}{\mathrm{d}t}. \tag{2-6}$$

由(2-1)式且注意到(2-2)式中 $u_0 > 0$，得到

$$\frac{\mathrm{d}u_i(t)}{\mathrm{d}x_i(t)} = \frac{\mathrm{d}[f_i^{-1}(x_i(t))]}{\mathrm{d}x_i(t)} > 0. \tag{2-7}$$

将(2-6)及(2-7)式代入(2-4)式，得到

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = -\sum_{i=1}^{N}\left[\frac{\mathrm{d}x_i(t)}{\mathrm{d}t}\right]^2\frac{\mathrm{d}[f_i^{-1}(x_i(t))]}{\mathrm{d}x_i(t)}. \tag{2-8}$$

由此式可以导出以下两点结论：

(1)　$E(t)$ 的导数恒非正且 $E(t)$ 有确下界，因此无论从什么初值出发以及无论何种输入，$E(t)$ 随着 $t \to \infty$ 将趋向一个极小点，在极小点上有 $\mathrm{d}E(t)/\mathrm{d}t = 0$. 若 $E(t)$ 有多个局部最小点，则所趋向的是哪个局部最小点取决于初值和输入.

(2) 在 $E(t)$ 的任一局部最小点上必然有 $\mathrm{d}x_i(t)/\mathrm{d}t = 0, i = 1,2,\cdots,N$，即每一神经元的输出保持恒定值. 系统有稳定解.

如果(2-2)式中的参数 u_0 充分小，则(2-3)式右侧第二项非常接近于0，这时能量函数 $E(t)$ 可取下列近似表示式：

$$E(t) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}x_i(t)x_j(t) - \sum_{i=1}^{N} I_i(0)x_i(t). \tag{2-9}$$

如果设 $\boldsymbol{X}(t) = (x_1(t), x_2(t), \cdots, x_N(t))$，$\boldsymbol{I}(0) = (I_1(0), I_2(0), \cdots, I_N(0))$，$\boldsymbol{W}$ 是元素为 w_{ij} 的($N \times N$)维对称方阵，则(2-9)式可以写成下列形式：

$$E(t) = -\frac{1}{2}\boldsymbol{X}(t)\boldsymbol{W}\boldsymbol{X}^{\mathrm{T}}(t) - \boldsymbol{I}(0)\boldsymbol{X}^{\mathrm{T}}(t). \tag{2-10}$$

2.2 连续时间 Hopfield 神经网络用于求解 TSP

TSP 是 Travelling Salesman Problem(旅行商问题)的缩写.其命题如下述.

设有 M 个城市,记为 A、B、C…,各城市之间的距离记为 d_{AB}、d_{AC}、d_{BC}….旅行商试图从任一城市出发走一条最短路径遍访所有其它城市后回到出发点,且每个城市只能访问一次.这样,问题归结为求一条通过所有城市且每个城市只通过一次的最短闭合路径.图 2-1 给出了 10 个城市 A、B、C、D、E、F、G、H、I、J 的例子,在此例中可以有 $\frac{1}{2}(10-1)! = 181\,440$ 种不同的闭合环路.例如,A E G D F I H B C J A 构成一个闭合环路,环路的路径长度为

$$d = d_{AE} + d_{EG} + \cdots + d_{JA}.$$

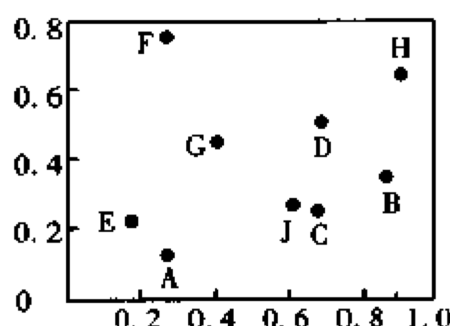


图 2-1

如果通过枚举法来选出最短闭合路径,计算量太大.下面介绍用 Hopfield 网络解此问题的方案.

为简单起见,讨论一个 $M = 5$ 的例子,设用 A、B、C、D、E 表示这 5 个城市,它们之间的距离为 d_{AB}、d_{AC}、….现在构造一个(5×5)矩阵,其行用 A、B、C、D、E 表示,其列用 1、2、3、4、5 表示.在表 2-1 给出的示例中,A 行 2 列元素为 1 而 A 行其它列元素为 0,表明 A 城市在环路中排第 2 位,余类推.这样,表 2-1 的矩阵表示环路:C A E B D C,环路路径长度为

$$d = d_{CA} + d_{AE} + d_{EB} + d_{BD} + d_{DC}.$$

表 2-1

	1	2	3	4	5
A	0	1	0	0	0
B	0	0	0	1	0
C	1	0	0	0	0
D	0	0	0	0	1
E	0	0	1	0	0

这个矩阵的特点是每行或每列中只有一个元素为 1,其他元素为 0;因此矩阵中只有 M 个元素为 1,其余皆为 0.

现在构造一个由($M \times M$)个神经元构成的 Hopfield 神经网络，其中每个神经元对应于上述表 2-1 矩阵中的一个元素，因此其编号用双重标记 mi 表示，m 表示行号，i 表示列号. 这样各神经元的输出、状态和输入将用双下标表示为 x_{mi}、u_{mi} 和 I_{mi}，而神经元之间的信号反馈权值可表示为 $w_{mi,nj}$(编号为 nj 神经元至编号为 mi 神经元之权). 为简便起见，用数字表示 M 个城市，即 m(或 n) $= 1,2,\cdots,M$，i(或 j) $= 1,2,\cdots,M$ 仍表示城市在环路中的排序. d_{mn} 表示序号为 m 和 n 的两个城市之间的距离. 用此网络求解 TSP 时首先应建立一个能量函数 E，当网络运行结束时网络的每行和每列应只有一个神经元的输出为 1，其他为 0，它们指示了一条闭合路径，其路径长度应是最短的或接近于最短的.

能量函数的计算公式如下(注意，变量 i 取模 M 数. 例如 $i = M + 1$ 时令 $i = 1$)

$$E(t) = \frac{\alpha}{2}\sum_{m=1}^{M}\sum_{\substack{i,j=1\\ i\neq j}}^{M}x_{mi}(t)x_{mj}(t) + \frac{\beta}{2}\sum_{i=1}^{M}\sum_{\substack{m,n=1\\ m\neq n}}^{M}x_{mi}(t)x_{ni}(t) + \frac{\zeta}{2}\Big[\sum_{m=1}^{M}\sum_{i=1}^{M}x_{mi}(t) - M\Big]^2 + \frac{\gamma}{2}\sum_{i=1}^{M}\sum_{\substack{m,n=1\\ m\neq n}}^{M}d_{mn}x_{mi}(t)[x_{n,i+1}(t) + x_{n,i-1}(t)]. \tag{2-11}$$

此式右侧第 1、2 两项的作用是在神经元阵列中每行和每列中至少有一个神经元的输出为 1 的条件下驱使其它各神经元的输出为 0；因为只有满足这一点才能使这两项达到其最小值，即 0 值. 第 3 项的作用是驱使阵列的($M \times M$)个神经元中有 M 个输出为 1，其它为 0；因为只有如此才可使此项达到其最小值 0. 第 3 项与前两项配合，使得这三项之和达到最小值时阵列的每行和每列中有且只有一个神经元的输出为 1 而其他神经元的输出皆为 0. 第 4 项给出由输出等于 1 的 M 个神经元连成的一条闭合路径的长度；当此项达到最小值时路径长度最短. α、β、ζ、γ 是 4 个大于 0 的控制参数.

如果将(2-3) 式改写为双下标形式并且暂时略去其右侧第二项的积分函数，可得

$$E(t) = -\frac{1}{2}\sum_{m,n=1}^{M}\sum_{i,j=1}^{M}w_{mi,nj}x_{mi}(t)x_{nj}(t) - \sum_{m=1}^{M}\sum_{i=1}^{M}I_{mi}(0)x_{mi}(t). \tag{2-12}$$

对比(2-11) 式和(2-12) 式，可得

$$\begin{cases} w_{mi,nj} = -\alpha\delta_{mn}(1-\delta_{ij}) - \beta\delta_{ij}(1-\delta_{mn}) - \zeta - \gamma d_{mn}(\delta_{j,i+1} + \delta_{j,i-1}), \\ I_{mi}(0) = \zeta M, \end{cases} \tag{2-13}$$

其中当 $i = j$ 时，$\delta_{ij} = 1$；当 $i \neq j$ 时，$\delta_{ij} = 0$. 将这一组权值用于(2-1) 式和(2-2) 式，且令各参数选值为 $\alpha = \beta = 500$，$\zeta = 200$，$\gamma = 500$. 令微分方程参数 $\tau = 1$，$u_0 = 0.02$，$I_{mi}(t) = I_{mi}(0)$. 以此来求解图 2-1 中 10 城市的例子(各城市间的距离按图中给定的尺度为准). 网络中各状态变量 u_{mi} 的初值 $u_{mi}(0)$ 按下列公式确定：

$$u_{mi}(0) = u_{00} + \delta u_{mi},\quad m = 1,2,\cdots,M,\quad i = 1,2,\cdots,M,$$

其中 δu_{mi} 是在$[-0.1u_0, 0.1u_0]$ 区间内均匀分布的随机变量，u_{00} 应使下式得到满

足：

$$\sum_{m=1}^{M}\sum_{i=1}^{M}x_{mi}(0) = 10.$$

注意其中 $x_{mi}(0) = f_{mi}(u_{mi}(0))$（见(2-2)式）且 $M = 10$. 在计算机上用数值计算方法可求解此微分方程组. 通过 20 组不同的初值所做的实验得到的结果表明，网络都能收敛，即在由 (10×10) 个神经元构成的矩阵中每行每列都只有一个输出为 1，而其他输出为 0. 因此每次运行都能确定一条闭合路径. 经检验，其长度都是最小的或接近于最小的. 例如，DHIFCEAJCB 和 DHIFEAGJCB 是其中的两个结果，前者的路径长度 $d = 2.71$，是最短路径，后者的路径长度 $d = 2.83$，接近于最短路径.

对不同的 TSP 课题所作的模拟实验研究表明，实验结果与参数 α、β、ζ、γ 和 u_0 的选值关系甚大，只有在恰当的范围内选择时才能得到有用的结果. 例如，u_0 选择太大时，不能使运行结果收敛到每行每列只有一个神经元输出为 1. 而其他输出为 0；u_0 太小时，所得的闭合路径距最优结果甚远. 在前 4 个参数中 γ 起关键作用.

TSP 是组合数学研究领域中著名的 NP 难题之一. 如采取枚举法逐个检验，其计算量按 $M!$ 的速度增加. 若采用 Hopfield 神经网络，只需作少量实验就可以得到最优或接近最优的结果，从而为解决此问题打开了一个新思路.

2.3 离散时间 Hopfield 神经网络

网络包含 N 个神经元，编号为 $i = 1,2,\cdots,N$. 每个神经元的输出即等于其状态，用 x_i 表示，x_i 只可能等于 1 或 -1. 令 $X = (x_1, x_2, \cdots, x_N)$，称为网络的状态(矢量). 各 x_i 和 X 按离散时间 k（k 为整数）而变化，因而可记之为 $x_i(k)$ 和 $X(k)$. 神经元 j 至神经元 i 的信号传送权为 w_{ij}，所有 w_{ij} 构成一个 $(N\times N)$ 权矩阵 W. W 是一个对称阵，即 $W^{T} = W$ 或 $w_{ij} = w_{ji}$. 权皆为实数，且满足 $\forall i, w_{ii}\geqslant 0$. 若给定网络的初始状态 $X(0)$，就可以按照下面给出的非同步方式或同步方式计算出 $k\geqslant 1$ 各时刻的状态 $X(k)$.

(1) 非同步运行方式(或称为异步方式或串行方式). 若已知 $X(k)$，则 $X(k+1)$ 可以按下列规则求得：按照轮流的方式或随机的方式从 $X(k)$ 中选出一个分量 $x_i(k)$，然后按照下列公式计算 $X(k+1)$ 的各个分量 $x_l(k+1)$，$l = 1,2,\cdots,N$：

$$\begin{cases} x_i(k+1) = \mathrm{sgn}\left[\sum_{j=1}^{N} w_{ij}x_j(k) - \theta_i\right]; \\ x_l(k+1) = x_l(k), \quad 当\ l\neq i, \end{cases} \tag{2-14}$$

其中 θ_i 是神经元 i 的阈值参数. $\mathrm{sgn}[\cdot]$ 是取符号函数，它可能采取下列两种形式之一.

第一种，设 $u(k) = \sum_{j=1}^{N} w_{ij}x_j(k) - \theta_i$，则

$$\mathrm{sgn}[u(k)] = \begin{cases} 1, & u(k)\geqslant 0; \\ -1, & u(k) < 0. \end{cases} \tag{2-15}$$

第二种，

$$\operatorname{sgn}[u(k)] = \begin{cases} 1, & u(k) > 0; \\ x_i(k), & u(k) = 0; \\ -1, & u(k) < 0. \end{cases} \tag{2-16}$$

(2) 同步运行方式(或称为并行方式). 由 $X(k)$ 计算 $X(k+1)$ 如下：

$$x_i(k+1) = \operatorname{sgn}\left[\sum_{j=1}^{N} w_{ij}x_j(k) - \theta_i\right], \quad i = 1,2,\cdots,N, \tag{2-17}$$

其中 sgn[·] 仍可采取(2-15) 式或(2-16) 式的形式.

网络的运行涉及两方面问题. 第一方面，网络共有 2^N 种不同的状态，是否存在若干个状态，当网络运行到这些状态时就不再改变. 这些状态相应于一个非线性动力系统的稳定态，称为**吸引子或定点**(其确切定义见下节所述部分). 下节将讨论吸引子的存在条件和有关定理. 第二方面，若网络有多个吸引子，那么从某个初始状态出发将最终收敛到哪个吸引子呢?这有关于网络的联想记忆功能和记忆容量的问题，将在下节讨论. 为简化起见下文中一律简称离散时间 Hopfield 神经网络为 Hopfield 神经网络.

2.4 有关 Hopfield 神经网络吸引子的基本定义和定理

2.4.1 吸引子的概念

定义 1 若网络状态 X 满足下列方程，则称其为网络的吸引子

$$X^{\mathrm{T}} = \operatorname{sgn}[WX^{\mathrm{T}} - \boldsymbol{\theta}^{\mathrm{T}}], \tag{2-18}$$

其中 sgn[U] 表示对 U 的每一个分量按(2-15) 式或(2-16) 式作 sgn[·] 运算后得到的矢量. 其次，$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$.

定理 1 如果按非同步方式运行，从任何初状态 $X(0)$ 出发，Hopfield 网络将最终收敛到一个吸引子.

证明 首先定义能量函数 $E(k)$，

$$E(k) = -\frac{1}{2}X(k)WX^{\mathrm{T}}(k) + X(k)\boldsymbol{\theta}^{\mathrm{T}}. \tag{2-19}$$

(1) 设 $\Delta E(k) = E(k+1) - E(k)$，则可以证明对任何 k，$\Delta E(k) \leqslant 0$ 成立. 设 $\Delta X(k) = X(k+1) - X(k)$，在异步运行时每一时刻 k 只有一个神经元的状态发生变化(设为第 i 神经元)，因此 $\Delta X(k)$ 中只有第 i 分量等于 $\Delta x_i(k)$，其他分量皆为0. 则可求得

$$\Delta E(k) = -\Delta x_i(k)\left[u_i(k) + \frac{1}{2}\Delta x_i(k)w_{ii}\right],$$

$$u_i(k) = \sum_{j=1}^{N} w_{ij}x_j(k) - \theta_i. \tag{2-20}$$

由(2-14) 式可知 $x_i(k+1) = \operatorname{sgn}[u_i(k)]$，这样，可分以下三种情况来讨论：

1) $x_i(k+1) = x_i(k), \Delta x_i(k) = 0$,则 $\Delta E(k) = 0$.

2) $x_i(k+1) = 1, x_i(k) = -1, \Delta x_i(k) = 2$. 由于 $x_i(k+1) = 1$,必有 $u_i(k) \geqslant 0$,已知 $w_{ii} \geqslant 0$,则由(2-20)式可得 $\Delta E(k) \leqslant 0$.

3) $x_i(k+1) = -1, x_i(k) = 1, \Delta x_i(k) = -2$. 由于 $x_i(k+1) = -1$,必有 $u_i(k) \leqslant 0$,已知 $w_{ii} \geqslant 0$,同样由(2-20)式可得 $\Delta E(k) \leqslant 0$.

因此,对任何 k,$\Delta E(k) \leqslant 0$ 得证.

(2) 易证明 $E(k)$ 有下确界.因此必然存在正整数 K,经过 K 步运行后将必然有 $\Delta E(K+k) = 0, \forall k \geqslant 0$. $K < \infty$,即达到 K 后,$E(k)$ 不再随 k 而变化.

(3) 可以分以下两种情况证明,只要 $E(k)$ 不再随 k 而变化,网络的状态即是吸引子或再经过有限多步运行即达到吸引子状态.

1) 若 $\Delta E(K+k) = 0$ 且 $\Delta x_i(K+k) = 0, i = 1,2,\cdots,N, \forall k \geqslant 0$ 成立,则 $X(K+k) = X(K), \forall k \geqslant 0$. 显然 $X(K)$ 即是吸引子.

2) 若 $\Delta E(K+k) = 0, \forall k \geqslant 0$,但是 $\Delta x_i(K+k) = 0$ 对某些 k 不成立. 如果某个 $\Delta x_i(K+k) \neq 0$,根据(2-20)式,必然有 $u_i(K+k) + \frac{1}{2}\Delta x_i(K+k) w_{ii} = 0$. 根据(1)中1)、3)可知,$u_i(K+k)$ 与 $\Delta x_i(K+k)$ 必同号,因此必须满足 $u_i(K+k) = 0$ 且 $w_{ii} = 0$. 若采取(2-16)式的 sgn[·]运算,当 $u_i(K+k) = 0$ 时,必然有 $\Delta x_i(K+k) = 0$,这说明原假设不成立. 若采取(2-15)式的 sgn[·]运算,当 $u_i(K+k) = 0$ 时若 $x_i(K+k) = -1$,则 $x_i(K+k+1) = 1$,这时 $\Delta x_i(K+k) = 2$. 显然,只要经过有限多步运行,将网络中所有服从上述规律的神经元的状态从 -1 转变为 1,则网络状态不可能再改变,即进入吸引子状态. 定理全部证完.

2.4.2 吸引子与能量函数局部最小点的关系

按照(2-19)式的定义,设能量函数 $E(k)$ 有一些局部最小点 X_ρ(见下述定义 2 和定义 3),则需回答下列正反两方面问题:在什么条件下局部最小点是吸引子?在什么条件下吸引子是局部最小点?在给出能量函数局部最小点定义后将分别回答这两个问题.

1. 能量函数局部最小点的定义

设 $\delta_i \boldsymbol{e}_i$ 是一个 N 维列矢量,其第 i 分量等于 δ_i,其他分量皆等于 0. 则可以给出局部最小点的以下两种定义.

定义 2 局部最小点 X_ρ 为满足下列条件者:设 X_ρ 的第 i 分量为 $x_{\rho i}$,令 $\delta_i = -2\text{sgn}[x_{\rho i}]$,则 $E(X_\rho) \leqslant E(X_\rho + \delta_i \boldsymbol{e}_i), i = 1,2,\cdots,N$,成立.

定义 3 称 X_ρ 为严格局部最小点,如果

$$E(X_\rho) < E(X_\rho + \delta_i \boldsymbol{e}_i), \quad i = 1,2,\cdots,N.$$

2. 能量函数局部最小点为吸引子条件.

定理 2 无论神经元函数 sgn[·] 取哪一种形式((2-15)式或(2-16)式),能量函数的严格局部极小点一定是吸引子;当 sgn[·]取(2-16)式形式时,能量函数的局部最小点是吸引子.

证明 若 X_ρ 是 E 的严格局部极小点,则由(2-20)式可以推出(设 $\Delta x_i(k) =$

δ_i)

$$\Delta E_i(k) = E(X_p + \delta_i e_i) - E(X_p) = -\delta_i[u_i(k) + \frac{1}{2}\delta_i w_{ii}] > 0.$$

为保证此式成立，δ_i 与方括弧中项应反符号. 若 $\delta_i = 2 > 0$，表明 $x_{pi} = -1$；为使 $\Delta x_i(k) = \delta_i = 2$，应保证 $u_i(k) \geq 0$；已知 $w_{ii} \geq 0$，这样，方括弧中项恒正，即不可能与 δ_i 反号. 结论是 $\Delta x_i(k) = \delta_i = 2$ 的假设不能成立，网络状态不可能改变. 若 $\delta_i = -2 < 0$，表明 $x_{pi} = 1$；为使 $\Delta x_i(k) = \delta_i = -2$，应保证 $u_i(k) < 0$；这样方括弧中项恒负，即不可与 δ_i 反号. 结论仍是 $\Delta x_i(k) = \delta_i = -2$ 的假设不成立，网络状态不变. 由于 i 可选任意值，所以 X_p 是吸引子.

若 X_p 是 E 的局部最小点，则只需讨论下列等式描述的情况（不等式情况与1同）：

$$\Delta E_i(k) = E(X_p + \delta_i e_i) - E(X_p) = -\delta_i[u_i(k) + \frac{1}{2}\delta_i w_{ii}] = 0.$$

若 $\delta_i = 2$，则 $x_{pi} = -1$；为保证上式成立，方括弧项应等于0，则必须有 $u_i(k) \leq 0$. 若 $u_i(k) < 0$，必有 $\Delta x_i(k) = \delta_i < 0$，与假设矛盾；若 $u_i(k) = 0$，必有 $\Delta x_i(k) = \delta_i = 0$，亦与假设矛盾. 故原假设不成立. 若 $\delta_i = -2$，则 $x_{pi} = 1$：为保证上式成立，必有 $u_i(k) \geq 0$. 若 $u_i(k) > 0$，必有 $\Delta x_i(k) = \delta_i > 0$，与假设矛盾；若 $u_i(k) = 0$，必有 $\Delta x_i(k) = \delta_i = 0$，亦与假设矛盾. 故原假设不成立. 证完.

3. 吸引子是能量函数局部极小点条件

定理3　若 $w_{ii} = 0, i = 1,2,\cdots,N$，则吸引子必为能量函数的局部极小点.

证明　设 X_r 是一个吸引子，由(2-20)式得

$$\Delta E_i = E(X_r + \delta_i e_i) - E(X_r) = -\delta_i[u_i + \frac{1}{2}\delta_i w_{ii}],$$

其中 $u_i = \sum_{j=1}^{N} w_{ij}x_{rj} - \theta_i$，$X_r = (x_{r1}, x_{r2}, \cdots, x_{rN})$，$\delta_i = -2\text{sgn}[x_{ri}]$. 由于 $w_{ii} = 0$，$\Delta E_i = -\delta_i u_i = 2\text{sgn}[x_{ri}]u_i = 2x_{ri}u_i$. 由于 X_r 是吸引子，x_{ri} 与 u_i 同号，因此对任何 i 皆有 $\Delta E_i \geq 0$. 所以 X_r 是局部最小点.

2.4.3　吸引子的其他性质

定理4　若 X_r 是吸引子，当 $w_{ii} = 0, i = 1,2,\cdots,N$，且 $\text{sgn}[0] = 1$ 时，$X_r + \delta_i e_i$，$i = 1,2,\cdots,N$，不是吸引子，其中 $\delta_i = -2\text{sgn}[x_{ri}]$.

证明　设 $X_r = (x_{r1}, x_{r2}, \cdots, x_{rN})$，$Y_r = X_r + \delta_i e_i = (y_{r1}, y_{r2}, \cdots, y_{rN})$；$x_{rj} = y_{rj}$，$j = 1,2,\cdots,N$ 且 $j \neq i$；$x_{ri} = -y_{ri}$. 由于 X_r 是吸引子且 $w_{ii} = 0$，则

$$x_{ri} = \text{sgn}[\sum_{\substack{j=1\\j\neq i}}^{N} w_{ij}x_{rj} - \theta_i] = \text{sgn}[\sum_{\substack{j=1\\j\neq i}}^{N} w_{ij}y_{rj} - \theta_i].$$

由此立即可导出 $y_{ri} = -\text{sgn}[\sum_{j=1}^{N} w_{ij}y_{rj} - \theta_i]$. 因此 Y_r 不是吸引子.

定理5　若 X_r 是吸引子，当 $\theta = 0$ 且按(2-16)式定义 $\text{sgn}[\cdot]$ 运算时，$-X_r$ 是

吸引子.

证明　当 sgn[·] 满足(2-16)式时，可作下列推导：由于 X_r 是吸引子，故 $X_r^T = \mathrm{sgn}[WX_r^T]$，则 $\mathrm{sgn}[W(-X_r^T)] = \mathrm{sgn}[-WX_r^T] = -\mathrm{sgn}[WX_r^T] = -X_r^T$. 所以 $-X_r$ 也是吸引子.

2.4.4　同步运行的情况

同步运行规则(2-17)式可以写成矢量形式

$$X^T(k+1) = \mathrm{sgn}[WX^T(k) - \theta^T]. \tag{2-21}$$

若 $X(k) = X_r$ 是一个吸引子，按照吸引子的定义((2-18)式)必然有 $X(k+1) = X(k)$，即按照同步方式运行时，网络状态一旦进入吸引子就不再改变. 如果从一个非吸引子状态出发，网络状态是否能收敛到一个吸引子呢？其条件由下列定理给出.

定理 6　若权矩阵 W 为非负定，且 sgn[·] 按照(2-16)式的规定运行，则网络从任何一个初始状态出发按同步方式运行时必收敛到一个吸引子.

证明　按照同步运行规则(2-21)式，可得

$$\Delta X(k) = X(k+1) - X(k).$$

引用(2-19)式，令 $\Delta E(k) = E(k+1) - E(k)$，经过简单推导可得

$$\Delta E(k) = -u(k)\Delta X^T(k) - \frac{1}{2}\Delta X(k) W \Delta X^T(k),$$

其中 $u^T(k) = WX^T(k) - \theta^T$. 在本章定理 1 的证明中，对(2-20)式所作的 1)、2) 两点分析表明：$u(k)$ 的第 i 分量 $u_i(k)$ 与 $\Delta X(k)$ 的相应分量 $\Delta x_i(k)$ 为同号($i = 1 \sim N$) 或者 $u_i(k) \cdot \Delta x_i(k) = 0$，因此 $u(k)\Delta X^T(k) \geqslant 0$. 若 W 为非负定，则 $\Delta X(k) W \Delta X^T(k) \geqslant 0$ 对任何 $\Delta X(k)$ 成立. 由于 $E(k)$ 有下确界，则通过有限多步运行网络的能量函数必然达到一个局部极小点. 根据定理 2 以及关于 sgn[·] 运算为(2-16)式的假定. 此局部最小点即是吸引子. 证毕.

如果 W 不满足非负定的要求，网络状态有可能陷入振荡. 举一个简单的例子，设有一个 $N = 2$ 的 Hopfield 网络，其 W 和 θ 如下所示：

$$W = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}, \quad \theta = (0,0).$$

若从初始态 $X(0) = (1,1)$ 出发运行，网络状态将在 $(1,1)$ 和 $(-1,-1)$ 之间振荡.

2.5　离散时间 Hopfield 神经网络的自联想记忆功能及其存储容量(记忆容量)

设有 M 个待记忆的 N 维二进矢量 $X^{(m)} = (x_1^{(m)}, x_2^{(m)}, \cdots, x_N^{(m)})$，$m = 1,2,\cdots,M$，其中 $x_i^{(m)} = 1$ 或 -1. 这些待记忆矢量可以用一个由 N 个神经元构成的离散时间 Hopfield 神经网络(简称 Hopfield 网络)来存储，其权矩阵 W 和阈值矢量 θ 是

$$W = \sum_{m=1}^{M}[(X^{(m)})^T X^{(m)} - I], \quad \theta = 0, \tag{2-22}$$

其中 I 是 $(N \times N)$ 维单位矩阵. 显然 W 是一个 $N \times N$ 对称阵且 $w_{ii} = 0, i = 1,2,\cdots,N$. W 是由诸记忆项 $X^{(m)}$ 的外积之和所构成. 此网络可作为一个联想记忆器,其功能是:若网络的初始状态 $X(0)$ 与某个记忆项 $X^{(m)}$ 比较接近(接近的定义见以下叙述),网络经过有限多步运行将达到并停留在状态 $X^{(m)}$, 就称由 $X(0)$ 通过此网络实现了对 $X^{(m)}$ 的**联想记忆**. 为有效地实现此功能,应该回答下列诸问题. 第一,各 $X^{(m)}$ 是否是吸引子?第二,是否存在以 $X^{(m)}$ 为中心的吸引域(当 $X(0)$ 在吸引域中时网络的状态将收敛到 $X^{(m)}$)?第三,网络能够记忆的矢量个数 M 受限于什么因素?与 N 有何关系?这些问题的解答还与各被存储项 $X^{(m)}$ 的性质有关. 例如,有些情况下各 $X^{(m)}$ 之间相互正交,即 $X^{(m_1)}(X^{(m_2)})^{\mathrm{T}} = 0, m_1 \neq m_2$. 但是,在更多的实际情况中 $X^{(m)}$ 为随机矢量,其各分量 $x_i^{(m)}$ 取 1 或 -1 的概率相等,且各个矢量之间和同一矢量各分量之间统计独立. 为了回答这些问题,下面首先给出一些基本定义,然后针对 $X^{(m)}$ 为随机矢量的情况给出用统计方法得到的结果.

定义 4　矢量 $X^{(m_1)}$ 和 $X^{(m_2)}$ 之间的汉明(Hamming)距离 $\mathrm{HD}(X^{(m_1)}, X^{(m_2)})$ 用下列公式计算:

$$\mathrm{HD}(X^{(m_1)}, X^{(m_2)}) = \frac{1}{2}\sum_{i=1}^{N} | x_i^{(m_1)} - x_i^{(m_2)} |. \tag{2-23}$$

它表示 $X^{(m_1)}$ 与 $X^{(m_2)}$ 二者互异分量的个数.

设在每一个记忆项 $X^{(m)}$ 周围有一个以其为中心的邻域,当网络初始状态 $X(0)$ 在此邻域中时,网络按同步方式((2-17) 式)或异步方式((2-14) 式)运行,最终收敛到 $X^{(m)}$ 的概率 p 非常接近于 1,则此邻域称为相应记忆项的**吸引域**. 吸引域可用以各 $X^{(m)}$ 为中心的球体来描述,球体表层的矢量与中心的汉明距离最大,越靠近中心时汉明距离越小. 设表层各矢量与中心的汉明距离为 $\rho N, 0 \leqslant \rho < \frac{1}{2}$, 则此球体称为吸引球,$\rho$ 称为吸引半径. ρ 是一个可供选择的参数,ρ 越大表示吸引域越大. 下面分同步一次运行直接吸引到中心,同步二次运行间接吸引到中心和异步运行吸引到中心三种情况,讨论 M、N、ρ、p 等参数之间的关系.

定理 7　若 Hopfield 网络由 N 个神经元构成,其权矩阵和阈值矢量满足(2-22)式,网络存储的记忆项为 $X^{(m)}, m = 1,2,\cdots,M$, 网络按(2-17) 式规定的同步方式运行且 $\mathrm{sgn}[\cdot]$ 运算符合(2-16) 式的规定. 设 ϵ 和 ρ 是两个固定参数,$0 < \epsilon \ll 1, 0 \leqslant \rho < \frac{1}{2}$, 则当 $X(0)$ 在以 ρN 为半径,$X^{(m)}$ 为中心的球体中时,若 M 满足(2-24) 式且 $N \to \infty$, 则 $X(1) = X^{(m)}$ 的概率 p 不小于 $\mathrm{e}^{-\epsilon}$.

$$M \leqslant (1-2\rho)^2 \frac{N}{2\ln N}\left[1 + \frac{\frac{1}{2}\ln(\ln N) + \ln(\epsilon\sqrt{4\pi})}{\ln N}\right]. \tag{2-24}$$

此式的误差量级为 $O(\frac{1}{\ln N})$, 详细证明可查阅文献 1.

定理 8　在本章定理 7 所述的约定下,当 $X(0)$ 在以 ρN 为半径,$X^{(m)}$ 为中心的球体中时,若 M 满足(1-25) 式且 $N \to \infty$, 则 $X(2) = X^{(m)}$ 的概率 p 不小于 $\mathrm{e}^{-\epsilon}$.

$$M \leqslant \frac{N}{2\ln N}\left[1+\frac{\frac{1}{2}\ln(\ln N)+\ln(\varepsilon\sqrt{4\pi})}{\ln N}\right]. \tag{2-25}$$

其证明可查阅文献 1. 可以看到，当 N 足够大从而使此式方括弧中的项接近于 1 时，Hopfield 网络的最大记忆容量接近于 $N/2\ln N$.

定理 7 的结论还可以用另一种方式表述：按照该定理所约定的条件，若给定记忆项的个数 M，则当 $X(0)$ 在以 ρN 为半径 $X^{(m)}$ 为中心的球体中时，$X(1)=X^{(m)}$ 的概率 p 由下式决定

$$p \approx \exp\left[-NQ\left(\sqrt{\frac{N}{M}}(1-2\rho)\right)\right]. \tag{2-26}$$

其中函数 $Q[\cdot]$ 由下式给出

$$Q(t)=\frac{1}{\sqrt{2\pi}}\int_t^{\infty}e^{-\frac{x^2}{2}}\mathrm{d}x.$$

这样，可以在理论上计算出一条 p 相对于 $\frac{M}{N}$ 而变化的曲线，当 $\frac{M}{N}$ 值较小时 p 非常接近于 1，当 $\frac{M}{N}$ 增大并超过某一值时，p 将急剧下降，很快趋向于 0. 当 $\rho \ll 0$ 时，此值约为 $\frac{1}{2\ln N}$，这相当于 $M=N/2\ln N$. 模拟实验结果表明，当 $N \geqslant 64$ 时，模拟结果与理论分析结果非常一致. 定理 8 也可以用类似的方式表述（见文献 1）.

当 Hopfield 网络按异步方式运行时，其结果简述如下. 在定理 7 所约定的条件下，若 $M=(1-\eta)N/2\ln N, 0 \leqslant \rho < (\frac{1}{2}-\delta)$，且 $0<\eta \ll 1, 0<\delta \ll 1$，则当 $X(0)$ 在一个以 $X^{(m)}$ 为中心且半径等于 $N\rho$ 的球体中时，经过有限次异步运行后收敛到 $X^{(m)}$ 的概率 p 非常接近于 1（参见文献[1]）. 这说明网络按异步方式运行时，其记忆项的最大存储容量也接近于 $N/2\ln N$.

2.6 双向联想记忆（BAM）网络

BAM 网络也称为异联想记忆网络，其结构如图 2-2 所示，其中上排包含 N 个

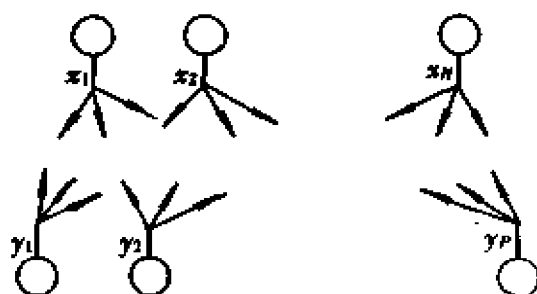


图 2-2

神经元，各神经元输出构成一个 N 维矢量 $X=(x_1,x_2,\cdots,x_N)$；下排包含 P 个神经

元,各输出构成一个 p 维矢量 $Y=(y_1,y_2,\cdots,y_P)$. X 和 Y 的各分量 x_i、y_j 只能等于 1 或 −1. 网络按照离散时间 k 运行,其运算规则如下

$$\begin{aligned}&Y(k+1)=\operatorname{sgn}[X(k)W-\theta],\quad X(k+1)=X(k),k=0,2,4,\cdots,\\&X(k+1)=\operatorname{sgn}[Y(k)W^{\mathrm T}-\mu],\quad Y(k+1)=Y(k),k=1,3,5,\cdots.\end{aligned}\tag{2-27}$$

其中 sgn[·] 的定义见(2-18)式,W 是一个 $N\times P$ 权矩阵,θ 是一个 P 维阈值行矢量,μ 是一个 N 维阈值行矢量. 若 θ 和 μ 皆为零矢量,则称为齐次 BAM 网络.

BAM 网络的能量函数 $E(k)$ 定义为

$$\begin{aligned}E(k)&=-X(k)W(Y(k))^{\mathrm T}+Y(k)\theta^{\mathrm T}+X(k)\mu^{\mathrm T}\\&=-Y(k)W^{\mathrm T}(X(k))^{\mathrm T}+Y(k)\theta^{\mathrm T}+X(k)\mu^{\mathrm T}.\end{aligned}\tag{2-28}$$

对于齐次 BAM 网络

$$E(k)=-X(k)W(Y(k))^{\mathrm T}=-Y(k)W^{\mathrm T}(X(k))^{\mathrm T}.\tag{2-29}$$

定理 9　若 BAM 从任意初始状态 $X(0)$、$Y(0)$ 出发按(2-27)式运行,则必然有 $\Delta E(k)=E(k+1)-E(k)\leqslant 0,k\geqslant 0$. 网络经过有限多步运行后必收敛到能量函数的一个局部极小点.

定理的证明参见文献 1,此处从略(其中 sgn[·] 运算需按照(2-16)式的规定).

齐次 BAM 网络的双向联想记忆功能如下所述:设有 M 对双向记忆矢量$(X^{(m)},Y^{(m)}),m=1,2,\cdots,M$,$X^{(m)}$ 为 N 维行矢量,$Y^{(m)}$ 为 P 维行矢量. 则可以构成一个齐次 BAM 网络,其权矩阵 W 是各对 $X^{(m)}$、$Y^{(m)}$ 的外积和

$$W=\sum_{m=1}^{M}(X^{(m)})^{\mathrm T}Y^{(m)}.\tag{2-30}$$

若网络的初始态 $X(0)$ 在 $X^{(m)}$ 的一个邻域内(例如以 $X^{(m)}$ 为中心,半径等于 ρN 的汉明球内,$0\leqslant\rho<\frac{1}{2}$),$Y(0)$ 为任意矢量,当网络经过有限多步运行后,其状态收敛到 $X^{(m)}$、$Y^{(m)}$,则称由 $X^{(m)}$ 实现了对 $Y^{(m)}$ 的联想. 类似地,也可以实现 $Y^{(m)}$ 对 $X^{(m)}$ 的联想. 虽然若干实例证实了这种功能(见文献 1),但是记忆项的存储容量有多大仍是一个待研究的问题.

3　自组织特征映射(SOFM)神经网络

SOFM 神经网络又称为 Kohonen 神经网络,它是一个由 P 个神经元构成的二维(平面)阵列,如图 3-1 所示. 每个神经元的输出用 y_i 表示,下标 i 按神经元在阵列中的位置逐行逐列编号. 网络的输入是 N 维行矢量 $X=(x_1,x_2,\cdots,x_N)$,X 的各分量 x_j 皆取实数值.

SOFM 网络实现 X 至 $Y=(y_1,y_2,\cdots,y_P)$ 的映射,对于任一特定的输入 X,输出 Y 的各分量中只有一个分量等于 1,设此分量为 y_l,而其它分量皆等于 0,l 称为该输入 X 的标号. 此映射功能的完成机制如下:网络中每个神经元 i(此编号即是其输出 y_i 的下标)有一个 N 维权矢量 $W_i=(w_{i1},w_{i2},\cdots,w_{iN})$,其各分量皆取实数值,则

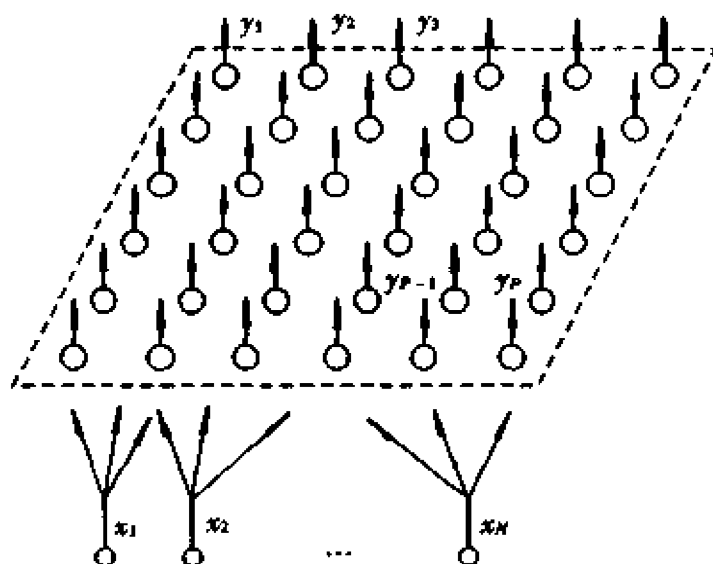


图 3-1

对于任一 X,其输出 Y 的各分量用下式计算:

$$l = \operatorname*{argmin}_{i} \| X - W_i \|,$$

$$y_i = \begin{cases} 1, & i = l; \\ 0, & i \neq l. \end{cases} \tag{3-1}$$

这样,SOFM 网络可以实现两类功能.第一类,作为矢量量化器(VQ).这时 l 表示中心为 W_l 的一个聚类区的标号.若输入 X 的标号为 l,表明 X 与 W_l 的欧氏距离小于 X 与其他 $W_i(i \neq l)$ 的欧氏距离.所以 l 可称为 VQ 标号.设有 M 个训练矢量 $X^{(m)}$,$m = 1,2,\cdots,M$,网络通过对它们的学习求得各 W_i 及相应的聚类区 Ω_i,$i = 1,2,\cdots,P$,每个 $X^{(m)}$ 属一个聚类区.SOFM 网络通过无监督的自组织学习所求得的各 W_i 应满足下列要求:

(1) 使得按下式计算的均方误差 J 达到最小,

$$J = \sum_{i=1}^{P} \sum_{X^{(m)} \in \Omega_i} \| X^{(m)} - W_i \|^2. \tag{3-2}$$

(2) 保持原来的拓扑特征,也就是当 $X^{(a)}$ 和 $X^{(b)}$ 两个矢量的欧氏距离较小时,其映射标号在二维阵列中处于较近的位置;反之,则较远.

(3) 保持原来概率密度分布的特征,即在概率密度函数 $p(X)$ 值较大的区域中分配有较多的神经元 i(W_i 在该区域中较密集);反之,神经元较少(W_i 较稀疏).第二类,作为分类器.这时 l 指示 X 所属的类别.SOFM 作为分类器时要求分类错误率达到最小.这时除了进行前述的自组织学习外,还要进行有监督的学习,即 LVQ.下面分别介绍实现这两类功能的 SOFM 的学习算法.

3.1 SOFM 用于 VQ 时的自组织学习算法

3.1.1 自组织学习算法框架

在给定训练集 $X^{(m)}, m = 1,2,\cdots,M$,以及网络神经元个数 P 的前提下,在图 3-2 给出的矩形或六角形神经元阵列形式中选择一种形式.然后按行对各神经元予以编号 i,每个神经元的权矢量为 W_i(W_i 和 $X^{(m)}$ 都是 N 维矢量).

自组织学习算法程序如下:

(1) 对 p 个权矢量赋随机初值 $W_i(0), i = 1,2,\cdots,P$.

(2) 设置最大迭代计算次数 K.

(3) 令迭代计算节拍 $k = 0$.

(4) 按序从训练集中取出一个矢量 $X^{(m)}$,计算 $\| X^{(m)} - W_i(k) \|, i = 1,2,\cdots, p$.求优胜编号 $l(k)$,

$$l(k) = \underset{i}{\operatorname{argmin}} \| X^{(m)} - W_i(k) \| .$$

(5) 在神经元阵列(见图 3-2) 中以 $l(k)$ 为中心划定一个邻域 $\Psi_l(k)$,此邻域随 k 的增大而逐渐缩小,所以将其写成 k 的函数.矩形形式的阵列中邻域按正方形扩大或缩小(图 3-2(a) 中虚线所示为每边包含 3 个神经元的正方形邻域).六角形形式的阵列中邻域按六边形扩大或缩小(图 3-2(b) 中虚线所示为边长等于 6 个神经元的六边形邻域).当 $k = 0$ 时,$\Psi_l(k)$ 应包括阵列中所有神经元;k 增大时 $\Psi_l(k)$ 缓慢缩小;当 $k = K$ 时 $\Psi_l(k)$ 缩为只包含神经元 $l(k)$ 本身.$\Psi_l(k)$ 可能取的几种形式将在下面 3.1.2 小节中给出.

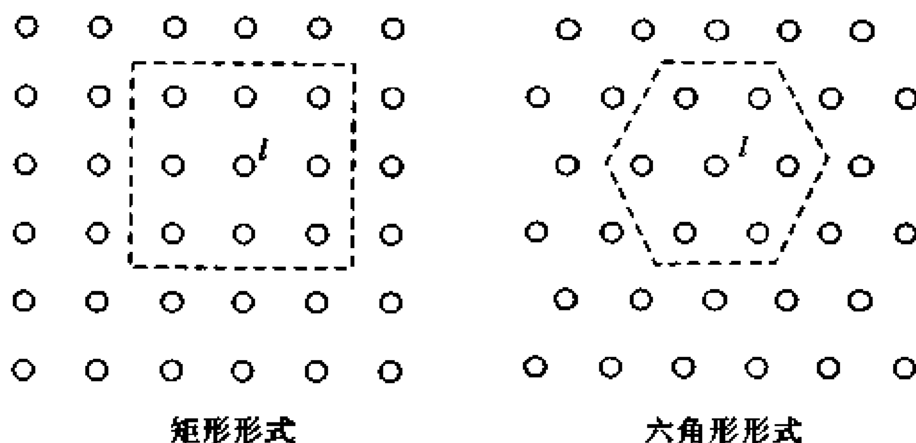


图 3-2

(6) 按下列公式计算新的权矢量

$$W_i(k+1) = \begin{cases} W_i(k) + \alpha(k)\beta(l,i,k)[X^{(m)} - W_i(k)], & i \in \Psi_l(k); \\ W_i(k), i \bar{\in} \Psi_l(k), \end{cases} \tag{3-3}$$

其中 $\alpha(k)>0$ 且随 k 的增大而缩小，称为步幅函数；$\beta(l,i,k)>0$，称为邻域函数，其中 l 表示 $l(k)$. 这两个函数可能采取的形式将在 3.1.2 小节给出.

(7) $k<K$?

若回答为是，令 $k=k+1$，转(4)；若回答为否，计算结束并输出计算结果：

$$W_i(K),\quad i=1,2,\cdots,P.$$

3.1.2　自组织学习算法中一些函数和参数的选择

1. $\Psi_l(k)$ 的选择

矩形阵列形式：$\Psi_l(k)$ 为正方形，其边长的神经元数为 $d=1,3,5,\cdots$. 设 D 是一个足够大的正奇数，使得无论 $l(0)$ 取何值时以 D 为边长的 $\Psi_l(0)$ 都能覆盖阵列中所有神经元. 则 $\Psi_l(k)$ 的边长 $d(k)$ 可按下列规律变化

$$d(k)=D-r,\quad \frac{r}{D}K\leqslant k<\frac{r+2}{D}K,$$
$$r=0,2,4,\cdots,D-1;$$

或

$$d(k)=D-r,\quad \frac{2^r}{2^D}K\leqslant k<\frac{2^{r+2}}{2^D}K,$$
$$r=0,2,4,\cdots,D-1.$$

六角形阵列形式：$\Psi_l(k)$ 为正六边形，其边长的神经元数 $d=1,2,3,\cdots$. 设 D 是一个使 $\Psi_l(0)$ 能覆盖全阵列的足够大正整数，则边长 $d(k)$ 可按下列规律变化

$$d(k)=D-r,\quad \frac{r}{D}K\leqslant k<\frac{r+1}{D}K,\quad r=0,1,2,\cdots,D-1;$$

或

$$d(k)=D-r,\quad \frac{2^r}{2^D}K\leqslant k<\frac{2^{r+1}}{2^D}K,\quad r=0,1,2,\cdots,D-1.$$

圆形邻域：$\Psi_l(k)$ 还可以为圆形. 如果设各行、各列相邻神经元的距离为 1，则可以以 $l(k)$ 为圆心，某个半径 $r_l(k)$ 所画出的圆来定义 $\Psi_l(k)$. 有关圆形邻域的半径 $r_l(k)$ 变化规律将在下面(3)中给出.

2. $\Psi_l(k)$ 为正方形或正六边形时 β 和 α 函数的选择

$$\beta(l,i,k)\equiv 1.$$

$\alpha(k)$ 可在下列几种形式中选择

$$\alpha(k)=\alpha_0/k,0<\alpha_0<1(\text{例如 }\alpha_0=0.5),$$

$$\alpha(k)=\alpha_0(1-\frac{k}{K}),\quad 0<\alpha_0<1,$$

$$\alpha(k)=\alpha_0 e^{-\eta\cdot\frac{k}{K}},\quad 0<\alpha_0<1,\eta>1.$$

此外，$\alpha(k)$ 可随着正方形或正六边形的缩小而同步缩小. 以正方形为例，$\alpha(k)$ 可按下式计算

$$\alpha(k)=(1-\frac{r}{D})\alpha_0,\quad \frac{r}{D}K\leqslant k<\frac{r+2}{D}K,$$

$$r = 0,2,4,\cdots,D-1, 0 < \alpha_0 < 1.$$

3. $\Psi_l(k)$ 为圆形域时 $r_l(k)$、β 和 α 的选择

选择 $r_l(0)$ 为足够大值使 $\Psi_l(0)$ 覆盖全阵列，然后可采取下列几种函数形式.

(1) $$r_l(k) = r_l(0) + [1 - r_l(0)]\frac{k}{K},$$
$$\beta(l,i,k) = 1,$$
$$\alpha(k) = (1 - \frac{k}{K})\alpha_0, \quad 0 < \alpha_0 < 1.$$

(2) $$r_l(k) = [r_l(0)]^{(1-\frac{k}{K})},$$
$$\beta(l,i,k) = 1,$$
$$\alpha(k) = \alpha_0 e^{-\eta\frac{k}{K}}, \quad 1 < \alpha_0 < 1, \eta > 1.$$

(3) $$r_l(k) = r_l(0),$$
$$\beta(l,i,k) = \exp\left\{-d_{li}\left[\frac{1.4}{r_l(0)} + \frac{k}{K}(5.6 - \frac{1.4}{r_l(0)})\right]\right\},$$
$$\alpha(k) \approx \alpha_0/k, \quad 0 < \alpha_0 < 1,$$

其中 d_{li} 是神经元 $l(k)$ 和神经元 i 之间的距离.

4. 神经元数 P、训练样本数 M 和叠代计算次数的选择

P 越大，则 VQ 精度越高（J 越小），而存储及计算量越大，在实用中一般根据 J 的要求确定 P. M 应比 P 大很多倍，以保证每个聚类区有足够多个训练样本 $X^{(m)}$（例如，每个聚类区平均有 30 ~ 100个样本，即 $M = 30P \sim 100P$）. K 应比 M 大若干倍（例如，$K = 20M$），以保证在迭代计算中充分利用每一个样本.

5. 初值设置

学习算法最终得到的均方误差 J（(3-2) 式）与初值关系甚大. 有很多好的初值设置方法，可查阅文献 2. 也可以设置多组初值，从不同的迭代计算结果中选一个最好的.

6. 再学习

在自组织学习结束后进行一轮再学习还能使 J 降低. 学习仍按 3.1.1 节框架进行，只是权调整计算公式（(3-3) 式）中的 $\Psi_l(k)$（邻域函数）只包括 $l(k)$ 本身一个神经元，$\beta(l,i,k) = 1, \alpha(k) = \alpha_{00}$（$\alpha_{00}$ 取甚小值，例如 $\alpha_{00} = 0.001$）. 此外，迭代次数 K 可选为一个足够大的数（例如 $K = 10\,000$）.

3.2　SOFM 用于模式识别时的学习算法(LVQ)

设输入矢量 X 分别属于 Q 种类别，则可以用一个具有 P 个神经元的 SOFM 网络对每一个输入的 X 进行分类，$P \geqslant Q$. 当 $P = Q$ 时，网络中每个神经元表示一种类别，神经元的编号即作为类别标号；当某个特定 X 输入时，若 $y_l = 1$ 且 $y_i = 0, i \neq l$，则判定 X 属于第 l 类. 当 $P > Q$ 时，一种类别可以由几个神经元表示，这些神经元中任何一个输出为 1 都指示 X 属于同一类别. P 的加大可以改善分类效果，却

加大了存储量和计算量.为简化起见,下文只讨论 $P = Q$ 的情况.

SOFM 网络用于模式识别时其学习过程分成两个阶段.设训练集为 $(X^{(m)}, l^{(m)})$, $m = 1,2,\cdots,M$,其中 $l^{(m)}$ 为 $X^{(m)}$ 所属的类别标号,M 为训练集的容量.第一阶段只用 $\{X^{(m)}\}$, $m = 1,2,\cdots,M$,对网络进行训练,采取 3.1.1 节所述的自组织学习算法.学习结束时每一个神经元周围形成一个聚类区,每个聚类区中包含的训练矢量 $X^{(m)}$ 可能具有不同标号 $l^{(m)}$,以一致标号最多者作为相应神经元的类别标号.第二阶段用 $(X^{(m)}, l^{(m)})$, $m = 1,2,\cdots,M$,对网络进行训练,这时采用有监督的学习算法.对较简单的分类问题可以用学习矢量量化(LVQ)算法,对于较复杂的分类问题(例如各类别所属 X 在值域中交迭较严重的情况),则应该采用 LVQ2 学习算法.

3.2.1 LVQ 算法

以第一阶段学习得到的权矢量作为第二阶段学习的初值 $W_i(0)$, $i = 1,2,\cdots,P$.按离散时序 $k = 0,1,2,\cdots$ 从训练集中取输入矢量 $X(k)$,其正确类别为 $\hat{l}(k)$(即训练集中规定类别).设网络按(3-1)式运行,判断 $X(k)$ 的类别为 $l(k)$,则可按下式由 $W_i(k)$ 求 $W_i(k+1)$,

$$W_i(k+1) = \begin{cases} W_i(k) + \alpha(k)\{X(k) - W_i(k)\}, & i = l(k) \text{ 且 } l(k) = \hat{l}(k), \\ W_i(k) - \alpha(k)\{X(k) - W_i(k)\}, & i = l(k) \text{ 且 } l(k) \neq \hat{l}(k), \\ W_i(k), & i \neq l(k), \end{cases} \tag{3-4}$$

其中 $\alpha(k)$ 可以取 3.1.2 节 2. 中给出的几种形式,只是 α_0 应选较小值(例如 $\alpha_0 = 0.05$).最大迭代计算次数为 K,K 应选为足够大的数值(例如,$K = 10M$).

3.2.2 LVQ2 算法

仍以第一阶段学习结果作为第二阶段学习初始权矢量,并且按时序 k 进行迭代计算.设 k 时刻从训练集中取出 $X(k)$ 送入网络,其正确类别为 $\hat{l}(k)$.设网络判断 $X(k)$ 的类别是 $l(k)$(第一优胜者),而第二优胜者为 $\gamma(k)$,即有

$$\| X(k) - W_{l}(k) \| < \| X(k) - W_{\gamma(k)} \| < \| X(k) - W_i \|,$$
$$i \neq l(k), i \neq \gamma(k).$$

再设有一个 $X(k)$ 的值域 $D(k)$,$D(k)$ 按下列不等式定义:

若 $\| X(k) - \frac{1}{2}(W_{l(k)} + W_{\gamma(k)}) \| \leqslant d$,则 $X(k) \in D(k)$,其中 d 是一个小正数.当 $X(k) \in D(k)$ 时,它与 $W_{l(k)}$ 的欧氏距离近似于它与 $W_{\gamma(k)}$ 的欧氏距离.这时各 W_i 可按下式进行迭代计算.

$$W_i(k+1)=\begin{cases} W_i(k),\ i=l(k)\ 且\ l(k)=\hat{l}(k), \\ W_i(k)+\alpha(k)\{X(k)-W_i(k)\}, \\ \quad i=\gamma(k)\ 且\ \gamma(k)=\hat{l}(k), X(k)\in D(k), \\ W_i(k)-\alpha(k)\{X(k)-W_i(k)\}, \\ \quad i=l(k)\ 且\ \gamma(k)=\hat{l}(k), X(k)\in D(k), \\ W_i(k), 所有其他情况, \end{cases} \tag{3-5}$$

其中 $\alpha(k)$ 仍可按 3.2.1 节 LVQ 算法中所述的原则选择. 这一算法的特点是, 只有当第一优胜者错误而第二优胜者正确且 $X(k)$ 位于此二者分界面附近时才对权矢量进行调整, 而其他情况一概不予调整. 实验结果表明, 对于许多复杂的分类问题, 这种算法优于模式识别中常用的理想贝叶斯分类器或 KNN 分类器(参见文献 2), 因此有较高的实用价值.

3.2.3　修正的 LVQ2 算法

在其他前提与 3.2.2 节所述 LVQ2 算法相同的条件下, 设 $l(k)$ 和 $\gamma(k)$ 分别为第一优胜神经元和第二优胜神经元的标号, 则可按下列公式进行各权矢量的迭代计算

(1) 若 $l(k)=\hat{l}(k)$,

$$W_i(k+1)=\begin{cases} W_i(k)+\alpha(k)\{X(k)-W_i(k)\}, & i=l(k), \\ W_i(k)-\alpha(k)\{X(k)-W_i(k)\}, & i=\gamma(k), \\ & 且\ |\,\|X(k)-W_{l(k)}\| \\ & -\|X(k)-W_{\gamma(k)}\|\,|\leqslant\theta, \\ W_i(k), & 所有其他情况, \end{cases} \tag{3-6}$$

其中 θ 是一个小正数.

(2) 若 $\gamma(k)=\hat{l}(k), l(k)\neq\hat{l}(k)$,

$$W_i(k+1)=\begin{cases} W_i(k)+\alpha(k)\{X(k)-W_i(k)\}, & i=\gamma(k), \\ W_i(k)-\alpha(k)\{X(k)-W_i(k)\}, & i=l(k), \\ W_i(k), & i\neq\gamma(k), i\neq l(k). \end{cases} \tag{3-7}$$

(3) 若 $\gamma(k)\neq\hat{l}(k), l(k)\neq\hat{l}(k)$,

$$W_i(k+1)=\begin{cases} W_i(k)+\alpha(k)\{X(k)-W_i(k)\}, & i=\hat{l}(k), \\ W_i(k), & i\neq\hat{l}(k). \end{cases} \tag{3-8}$$

此式中的 $\alpha(k)$ 仍可按 3.2.1 节的 LVQ 算法进行. 修正 LVQ2 算法较 LVQ2 算法简单, 二者所实现的网络性能相接近.

参 考 文 献

1 杨行峻,郑君里编著.人工神经网络.北京:高等教育出版社,1992.
2 杨行峻,迟惠生等编著.语音信号数字处理.北京:电子工业出版社,1995.
3 Rumelhart D E, McClelland J L. Parallel Distributed Processing, MIT Press, Cambridge, MA, 1986, Vol.1 ~ Vol.2.
4 Hopfield J J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proc. Natl. Acad. Sci. U.S.A., 1982, Vol.79, 2554 ~ 2558.
5 Vapnik V N. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

第 20 篇

模糊数学

编　者　王国俊
审校者　张文修

目　　录

引　　言

人类是依靠概念进行逻辑思维的，通常希望能通过概念去刻画事物的本质属性，而将该类事物的全体清楚地界定出来，也就是希望能通过概念的内涵而清楚地确定其外延．经典集合论在这方面取得了巨大的成功，现代数学的各个分支无一不是建立在经典集合论的基础上的．然而，现实生活乃至科学技术领域中有大量的事物是难以通过精确地刻画其内涵而界定其外延的．另一方面，对于复杂的大系统而言，精确性与复杂度往往形成尖锐的矛盾，为了将其复杂度降低到可以实际操作的范围之内，就不得不牺牲一些精确性．基于这种情况，美国加州大学贝克莱分校著名的控制论专家扎德(L.A.Zadeh)教授于1965年提出了模糊集的概念①，用以表示界限不明确的那类事物．模糊集概念一经提出，立即受到科学技术界的广泛关注．30多年来，基于模糊集概念的理论学科与应用学科纷纷建立，并得到了迅速的发展，其中模糊系统方法在工业领域的成功应用尤其令人瞩目．

从纯数学的角度看，一个模糊集是从一个普通集 X 到[0,1]的映射，是 X 上特征函数概念的推广，自然也就是 X 的子集概念的推广．把这种推广用于数学的各个分支就可得到相应的模糊数学的各个分支．由于仅取0与1两个值的特征函数被连续取值于[0,1]的模糊集所取代，所以模糊数学的各分支一般都比相应的原数学分支复杂，且针对这种较复杂的情况许多新的方法被提出．然而更为重要的是模糊集概念被得到应用，特别是在模糊控制上得到应用．这时与其说是模糊集方法，不如说是模糊集思想取得了普遍的成功．本篇仅介绍模糊数学的一些理论分支，以及模糊集理论的应用．由于篇幅所限，在介绍模糊数学的某个分支(如模糊代数、模糊拓扑等)时是假定了读者熟悉与该分支相应的经典数学分支的．同时仅限于给出能反映该分支的基本思想的理论框架．

1　模糊集与映射

1.1　模糊集及其运算

1.1.1　模糊集

按照康托尔(G.Cantor)集论的观点，给定一个性质 P 就可得出一个集 A，它由

① Zadeh L A. Fuzzy sets. Inform. and Control, 1965(8): 338 ~ 353

一切具有性质 P 的对象 x 组成. 这时 A 可写作 $A=\{x:x$ 具有性质 $P\}$. A 是性质 P 的外延,性质 P 是 A 的内涵,所以康托尔集论是以承认事物的内涵可以清楚地确定其外延为前提的. 通常把考虑的对象限制在某个集 X 之内,把集 X 称为论域,然后利用性质 P 去刻画出 X 的子集来. 这个子集可以写作

$$A=\{x\in X:x\text{ 具有性质 }P\}. \tag{1-1}$$

比如,取 X 为实数集 $\mathbf{R}$,性质 P 为"平方小于9",则 P 刻画出 $\mathbf{R}$ 的一个子集 A,$A=\{x\in\mathbf{R}:x^2<9\}=(-3,3)$. 这时对于论域 $\mathbf{R}$ 中的每个 x,x 是否属于 A 是完全清楚的,从而 A 可以用一个定义在集 X 上且取1与0两个值的函数来表示. 若用 $A:X\to\{0,1\}$ 表示这个函数,则有

$$A(x)=\begin{cases}1 & (x\in A),\\ 0 & (x\in X-A).\end{cases} \tag{1-2}$$

这个函数叫做 X 的子集 A 的**特征函数**.

然而有许多事物其内涵以及相应的外延是不清楚的. 比如,用 X 表示全体中国人构成的集,性质 P 为"年轻",则(1-1)式中的 A 是不清楚的,因为"年轻"的含义是不清楚的. 当然,可以把性质 P 具体化,比如,规定年龄在18岁到40岁之间为年轻,这样(1-1)式中的 A 就是明确的了. 但这又与人们心目中"年轻"的含义不尽相同,因为把一个41岁的健壮的人与一个39岁的体弱多病的人相比较,人们会认为前者更年轻一些. 可见"年轻"这个性质是不能刻画一个明确的集合的. 如果把性质 P 定为"足智多谋",则(1-1)式中的 A 就更加不清楚. 事实上人们进行思维时除了用到内涵与外延都清楚的概念以外,更大量运用的是内涵与外延都不清楚的概念. 同时不只是在日常生活中,在科学技术领域中也经常要和这种不清楚的概念所描述的事物打交道. 比如,要制造一种飞行速度快、灵活且不易被敌方雷达发现的侦察机就涉及"快"、"灵活"和"不易被发现"这些不清楚的性质. 由这类概念所描述的事物是难以用经典的康托尔集来表示的. 许多学者对此早有看法. 1965年美国加州大学贝克莱分校著名的控制论专家扎德教授首次正式提出了模糊集的概念,用以表示上述种种界限不清的事物.

定义1　设 X 是非空集,X 的**模糊子集** A 是一个函数 $A:X\to[0,1]$. $\forall x\in X$,称 $A(x)$ 为 x 对 A 的隶属度,称 X 为 A 的**论域**. 这时也说,A 是论域 X 上的**模糊集**,简称为 F 集,称 $\mathrm{supp}A=\{x\in X:A(x)>0\}$ 为 A 的承集. X 上模糊集的全体记作 $\mathscr{F}(X)$.

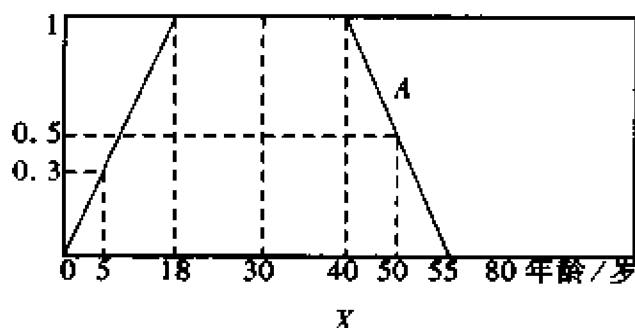


图 1-1

例1　年轻人的全体可以用一个模糊集 A 来表示. 如图1-1所示,用横轴上的一段 X 表示不同年龄的人. 用纵轴上的[0,1]表示隶属度. 如果一个人 x 的年龄为30岁,由图看出,x 对 A 的隶属度为

$$A(x)=1,$$

那么 x 是名副其实的年轻人. 如果一个人 y 的年龄为50岁,由图看出,

$$A(y) = 0.5,$$

那么 y 只在0.5的程度上算是年轻人. 如果 z 是一个5岁的儿童,则 z 也不在通常所说的年轻人之列,由图看出,z 是年轻人的隶属度为0.3. 这里"年轻"的内涵是不清晰的,相应地其外延也就不清晰. 但一个80岁的老人绝不是年轻人,他相应于 A 的隶属度等于0.

图1-1中曲线 A 从40岁开始下降,从18岁往前也开始下降. 这两个年轻界限不是不可改变的. 不同的人可能采取不同的界限. 同时,下降的曲线也不必取为直线,也可取为二次曲线,乃至对数或指数曲线等. 可见模糊集在来源上有不确定性.

当映射 A 只取1与0两个值时,A 就是 X 的某子集的特征函数,仍用 A 记这个子集,则可得(1-2)式. 可见,模糊集是康托尔集概念的扩充. 为与模糊集相区别,以下把经典的康托尔集称为分明集. 不加定语"模糊"的集均指分明集.

定义2 设 $A \in \mathscr{F}(X)$,即 A 是论域 X 上的模糊集,$x \in X, \lambda \in (0,1]$. 设映射 $x_\lambda: X \to [0,1]$ 为

$$x_\lambda(t) = \begin{cases} \lambda & (t = x), \\ 0 & (t \neq x). \end{cases} \tag{1-3}$$

称 x_λ 为 X 上的**模糊点**,简称为 F 点. λ 叫 x_λ 的高,x 叫 x_λ 的承点. 当 $\lambda \leqslant A(x)$ 时,称 x_λ 属于 A,记作 $x_\lambda \in A$.

显然,模糊点就是承集为单元素集的模糊集.

1.1.2 模糊集的运算

因为 X 上的模糊集 A 实际上是从 X 到 $[0,1]$ 的函数,所以模糊集之间的序关系和运算就是相应的函数间的序关系和运算.

定义3 设 $A, B \in \mathscr{F}(X), \forall i \in I, A_i \in \mathscr{F}(X)$.

1° 称 A 包含于 B,记作 $A \leqslant B$,若 A 与 B 作为函数满足 $A \leqslant B$,即 $\forall x \in X, A(x) \leqslant B(x)$. 这时也说 B 包含 A.

2° $\{A_i\}$ 的并记作 $\bigvee_{i \in I} A_i$,交记作 $\bigwedge_{i \in I} A_i$,分别定义为

$$(\bigvee_{i \in I} A_i)(x) = \sup_{i \in I} A_i(x),$$
$$(\bigwedge_{i \in I} A_i)(x) = \inf_{i \in I} A_i(x) \quad (x \in X).$$

3° 设 $\lambda \in [0,1]$. 用 λA 表示 $[\lambda] \wedge A$,这里 $[\lambda]$ 是在 X 上取常值 λ 的函数.

4° $A' = 1 - A$ 叫 A 的**伪补**,定义为

$$A'(x) = 1 - A(x) \quad (x \in X).$$

因为 $[0,1]$ 是完全分配格,所以 $\mathscr{F}(X) = [0,1]^X$ 按上述包含序(即乘积序)构成一完全分配格,且有最大元 $[1]$ 与最小元 $[0]$. 它们也分别简写为1与0,或 X 与 $\varnothing$,或 1_X 与 0_X. $\mathscr{F}(X)$ 上还有一个**逆序对合对应**"′",即

$$A \leqslant B \quad 当且仅当 \quad B' \leqslant A',$$
$$(A')' = A. \tag{1-4}$$

因为从逆序对合对应的存在可推得德·摩根(De Morgan)对偶律,所以下述命题成立:

命题1　$\mathscr{F}(X)$ 是具有逆序对合对应的完全分配格. 设

$$A_i \in \mathscr{F}(X) \quad (i \in I),$$

则德·摩根对偶律成立,即

$$\begin{cases}(\bigvee_{i \in I} A_i)' = \bigwedge_{i \in I} A_i', \\ (\bigwedge_{i \in I} A_i)' = \bigvee_{i \in I} A_i'.\end{cases} \tag{1-5}$$

注1　许多文献中用 $\underset{\sim}{A}$, $\underset{\sim}{B}$ 等表示模糊集,或者把模糊集与表示它的函数分开,再引入所谓隶属函数 $\mu_{\underset{\sim}{A}}$, $\mu_{\underset{\sim}{B}}$ 等. 在论域 X 是有限集 $\{x_1, x_2, \cdots, x_n\}$,或无限集的情形,更有用

$$\underset{\sim}{A} = \frac{\mu_{\underset{\sim}{A}}(x_1)}{x_1} + \frac{\mu_{\underset{\sim}{A}}(x_2)}{x_2} + \cdots + \frac{\mu_{\underset{\sim}{A}}(x_n)}{x_n} = \sum_{i=1}^{n} \frac{\mu_{\underset{\sim}{A}}(x_i)}{x_i} \tag{1-6}$$

或

$$\underset{\sim}{A} = \int_{x \in X} \frac{\mu_{\underset{\sim}{A}}(x)}{x} \mathrm{d}x \tag{1-7}$$

来表示模糊集者. 为简明计,这里不采用这些记法.

1.1.3　模糊集的分解

X 上仅取两个值0与 λ 的模糊集是简单的,其图像如图1-2所示. X 上的任何一个模糊集都可用这种简单模糊集的并来表示. 为此先引入一个定义.

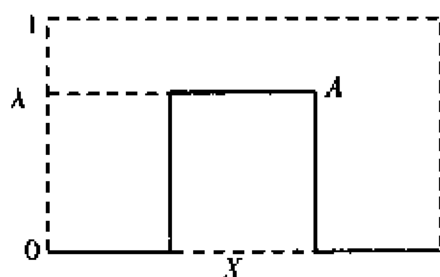


图 1-2

定义4　设 $A \in \mathscr{F}(X)$, $r \in [0,1]$,则

1° 称 $A_r = \{x \in X: A(x) \geqslant r\}$ 为 A 的 **r 截集**.

2° 称 $A_{(r)} = \{x \in X: A(x) > r\}$ 为 A 的 **r 强截集**.

显然, A 的承集就是 A 的 0 强截集 $A_{(0)}$.

r 截集与 r 强截集都是 X 的分明子集. 由于对 X 的分明子集和它的特征函数不需加以区别,因此, X 的分明子集也可看做是 X 上的模糊集. 它们是 X 上仅取 0 与 1 两个值的简单模糊集.

推论1　设 A 是简单模糊集, $r \in [0,1]$,则 $rA = [r] \wedge A$ 自然仍是简单模糊集.

定理1 (分解定理)　设 $A \in \mathscr{F}(X)$, $\mathbf{Q}$ 是有理数集,则

$$A = \bigvee_{r \in [0,1]} rA_r = \bigvee_{r \in \mathbf{Q} \cap [0,1]} rA_r = \bigvee_{r \in [0,1]} rA_{(r)} = \bigvee_{r \in \mathbf{Q} \cap [0,1]} rA_{(r)}.$$

关于 A 的截集与分解定理可分别参看图 1-3 和图 1-4.

1.1.4　模糊集的模糊性度量

模糊集是将特征函数的取值从1与0开拓到[0,1]而得的. 如果分别把1与0作为"是"与"非"的标准,则与1或0比较接近的数就具有比较清楚的是或非的性质,或者说它们的模糊性比较小. 相反地,远离0与1的数模糊性就大一些,其中模糊性

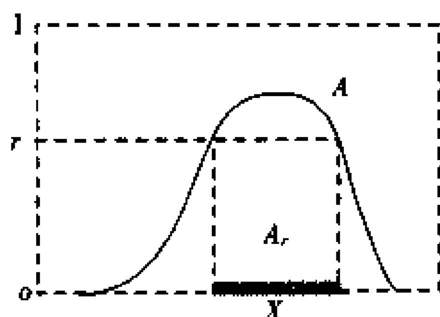


图 1-3

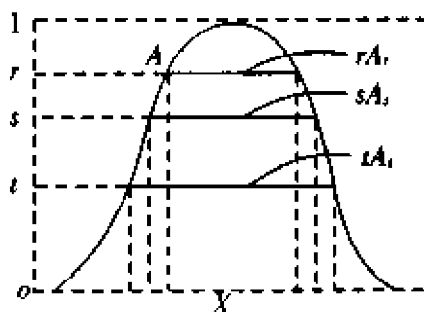


图 1-4

最大者当属值为 1/2 的了.由此有如下定义.

定义 5 设映射 $d:\mathscr{F}(X)\to[0,1]$ 满足条件:

1° $d(A)=0$,当且仅当 $A\in\mathscr{P}(X)$,即 A 为 X 的分明子集,

2° $A(x)$ 恒等于$\frac{1}{2}$ 时 $d(A)=1$,

3° 当 $\forall x\in X$,$\left|A(x)-\frac{1}{2}\right|\leqslant\left|B(x)-\frac{1}{2}\right|$时 $d(A)\geqslant d(B)$,

则称 d 为 $\mathscr{F}(X)$ 上的**模糊度**,或称 $d(A)$ 为 A 的模糊度.

由上述条件 3° 立即得出 $d(A')=d(A)$,因为容易证明

$$\left|A'(x)-\frac{1}{2}\right|=\left|A(x)-\frac{1}{2}\right|$$

对于一切 $x\in X$ 都成立.

例 2 设 X 是有限集,$X=\{x_1,x_2,\cdots,x_n\}$.令

$$H(A)=-\frac{1}{n\ln 2}\sum_{i=1}^{n}\Big(A(x_i)\ln A(x_i)+A'(x_i)\ln A'(x_i)\Big),\tag{1-8}$$

易验证 $H:\mathscr{F}(X)\to[0,1]$ 满足定义 5 的各条件,所以 H 是模糊度,称为申农(C.E. Shannon)的**模糊熵**.

例 3 设 $X=\{x_1,x_2,\cdots,x_n\}$.令

$$K_1(A)=\frac{2}{n}\sum_{i=1}^{n}|A(x_i)-A_{1/2}(x_i)|,\tag{1-9}$$

这里 $A_{1/2}$ 是 A 的$\frac{1}{2}$ 截集,$K_1:\mathscr{F}(X)\to[0,1]$ 满足定义 5 的各条件,所以 K_1 是模糊度,称为**模糊指标**.再令

$$K_2(A)=\frac{2}{\sqrt{n}}\Big(\sum_{i=1}^{n}(A(x_i)-A_{1/2}(x_i))^2\Big)^{1/2},$$

则 K_2 也是模糊度.

由以上各例可知,满足定义 5 的模糊度可以是多种多样的.

1.1.5 模糊集间的距离与贴近度

X 上的模糊集就是从 X 到$[0,1]$ 的函数,所以凡是关于度量函数间距离的方法

都可以用来度量模糊集之间的距离.但描述模糊集之间接近程度的方法不限于距离方法,还有**贴近度**方法.

定义6 设 $M_p:\mathscr{F}(X)\times\mathscr{F}(X)\to[0,+\infty]$ 满足

1° 当 $X=\{x_1,x_2,\cdots,x_n\}$ 时,

$$M_p(A,B)=\left(\sum_{i=1}^{n}|A(x_i)-B(x_i)|^p\right)^{1/p}, \tag{1-10}$$

2° 当 $X=[a,b]$ 时,

$$M_p(A,B)=\left(\int_a^b|A(x)-B(x)|^p\mathrm{d}x\right)^{1/p}, \tag{1-11}$$

则 M_p 是 $\mathscr{F}(X)$ 上的距离,称为**闵可夫斯基(H.Minkowski)距离**.

例4 在(1-10)式与(1-11)式中令 $p=1$,所得距离称为**汉明(R.W.Hamming)距离**;令 $p=2$,所得距离就是通常的**欧几里德(Euclidean)距离**.还可将闵可夫斯基距离加以改造,使距离的值不超出[0,1]的范围.令

$$M'_p(A,B)=\left(\frac{1}{n}\sum_{i=1}^{n}|A(x_i)-B(x_i)|^p\right)^{1/p}, \tag{1-12}$$

或

$$M'_p(A,B)=\left(\frac{1}{b-a}\int_a^b|A(x)-B(x)|^p\mathrm{d}x\right)^{1/p}, \tag{1-13}$$

所得距离的值就不超出[0,1],称为**相对闵可夫斯基距离**.

定义7 设 $\omega:X\to[0,1]$ 满足

1° 当 $X=\{x_1,x_2,\cdots,x_n\}$ 时,$\sum_{i=1}^{n}\omega(x_i)=1$,

2° 当 $X=[a,b]$ 时,$\int_a^b\omega(x)\mathrm{d}x=1$,

则称 ω 为**权函数**.定义 $M_\omega:\mathscr{F}(X)\times\mathscr{F}(X)\to[0,1]$ 如下:

1° 当 $X=\{x_1,x_2,\cdots,x_n\}$ 时,

$$M_\omega(A,B)=\sum_{i=1}^{n}\omega(x_i)|A(x_i)-B(x_i)|; \tag{1-14}$$

2° 当 $X=[a,b]$ 时,

$$M_\omega(A,B)=\int_a^b\omega(x)|A(x)-B(x)|\mathrm{d}x, \tag{1-15}$$

则 M_ω 是 $\mathscr{F}(X)$ 上的距离,称为**加权汉明距离**.

定义8 设 (E,ρ) 为度量空间,A 与 B 是 E 的非空子集.令

$$\rho_H^*(A,B)=\sup\{\rho(x,B):x\in A\}=\sup\{\inf_{y\in B}\rho(x,y):x\in A\}, \tag{1-16}$$

$$\rho_H(A,B)=\max\{\rho_H^*(A,B),\rho_H^*(B,A)\},$$

则 ρ_H 为 $\mathscr{P}(E)-\varnothing$ 上的伪距离,即以下条件成立.

1° 当 $A=B$ 时,$\rho_H(A,B)=0$,

2° $\rho_H(A,B)=\rho_H(B,A)$,

3° $\rho_H(A,C)\leqslant\rho_H(A,B)+\rho_H(B,C)$.

称 ρ_H 为 $\mathscr{P}(X)-\varnothing$ 上的**豪斯多夫(Hausdorff)伪距离**,其中 $\mathscr{P}(X)$ 表示 X 的一切子集构成的集,$\varnothing$ 是空集.

注意,当 $\rho_H(A,B)=0$ 时,未必有 $A=B$,但可证 A 与 B 的闭包相等.例如,当 E 为实直线时,令 $A=(a,b)$,$B=[a,b]$,则 $\rho_H(A,B)=0$,但 $A\neq B$.

定义 9 令

$$\mathscr{F}^*(X)=\{A\in\mathscr{F}(X):A\text{ 是从 }X\text{ 到}[0,1]\text{ 上的满射}\},$$

设 $A,B\in\mathscr{F}^*(X)$,令

$$\rho(A,B)=\sup\{\rho_H(A_\lambda,B_\lambda):\lambda\in[0,1]\},\tag{1-17}$$

则由 A,B 为满射可知,A_λ 与 B_λ 总是非空的,从而 $\rho_H(A_\lambda,B_\lambda)$ 存在.ρ 为 $\mathscr{F}^*(X)$ 上的伪距离,称为 $\mathscr{F}^*(X)$ 上的豪斯多夫伪距离.

例 5 设 $X=R$,$A,B\in\mathscr{F}(X)$,A 在区间 $[0,1]$ 上取值 1,在 $[1,2]$ 上是斜率为 -1 的线段,在其他处值为零;B 在 $[1,2]$ 与 $[2,3]$ 上分别是斜率为 1 和 -1 的线段,在其他处值为零(见图 1-5).这时

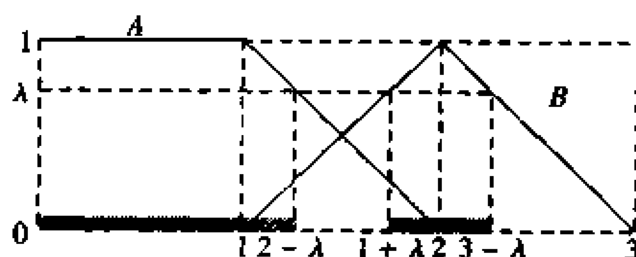


图 1-5

$$A_\lambda=[0,2-\lambda],\quad B_\lambda=[1+\lambda,3-\lambda].$$

所以 A_λ 与 B_λ 间的豪斯多夫伪距离为

$$\rho_H(A,B)=\max\{\rho_H^*(A,B),\rho_H^*(B,A)\}=\max\{1+\lambda,1\}=1+\lambda.$$

而 A 与 B 间的豪斯多夫伪距离为

$$\rho(A,B)=\sup\{\rho_H(A_\lambda,B_\lambda):\lambda\in[0,1]\}=2.$$

定义 10 设映射 $N:\mathscr{F}(X)\times\mathscr{F}(X)\to[0,1]$,若

1° $N(A,A)=1$,

2° $N(A,B)=N(B,A)$,

3° 当 $\forall x\in X$,$|A(x)-C(x)|\geqslant|A(x)-B(x)|$ 时,

$$N(A,C)\leqslant N(A,B),$$

则称 N 为 $\mathscr{F}(X)$ 上的**贴近度**.

定义 11 若分别在 $X=\{x_1,x_2,\cdots,x_n\}$ 与 $X=[a,b]$ 的情形,有

$$N_H(A,B)=1-\frac{1}{n}\sum_{i=1}^{n}|A(x_i)-B(x_i)|$$

与

$$N_H(A,B)=1-\frac{1}{b-a}\int_a^b|A(x)-B(x)|\,dx,$$

则 N_H 是贴近度,称为**汉明贴近度**.

若

$$N_E(A,B)=1-\frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n}(A(x_i)-B(x_i))^2\Big)^{1/2},$$

或

$$N_E(A,B)=1-\frac{1}{\sqrt{b-a}}\left(\int_a^b(A(x)-B(x))^2\mathrm{d}x\right)^{1/2},$$

则 N_E 为贴近度,称为**欧几里德(简称欧氏)贴近度**.

在定义11中贴近度都是用1减去某个距离函数而得到的.一般来说,设 ρ 是 $\mathscr{F}(X)$ 上的取值不超出[0,1]的伪距离,则 $1-\rho$ 就是贴近度.这里假定两个模糊集 A 与 B 是函数,它们越接近,$\rho(A,B)$ 就越小.

例6 贴近度与模糊度有某种联系.设 d 是 $\mathscr{F}(X)$ 上的模糊度,则由

$$\overline{N}(A,B)=d(A\ominus B)$$

确定的映射 $N:\mathscr{F}(X)\times\mathscr{F}(X)\to[0,1]$ 就是 $\mathscr{F}(X)$ 上的贴近度.这里 $A\ominus B\in\mathscr{F}(X)$,其定义为

$$(A\ominus B)(x)=\frac{1}{2}(1+|A(x)-B(x)|)\quad(x\in X).\qquad(1\text{-}18)$$

比如,取模糊度 d 为(1-9)式中的 K_1,注意 $(A\ominus B)(x)\geqslant\frac{1}{2}$ 恒成立,从而 $(A\ominus B)_{\frac{1}{2}}$ 恒等于1,则易验证由 $d(A\ominus B)$ 确定的贴近度就是汉明贴近度.

1.1.6 不同论域上的模糊集的乘积

不同论域上的模糊集之间可以相乘.

定义12 设 X_i 是非空分明集($i\in I$),$X=\prod\limits_{i\in I}X_i$ 是各 X_i 的乘积.设

$$A_i\in\mathscr{F}(X_i)\quad(i\in I),$$

则各 A_i 的乘积 $A=\prod\limits_{i\in I}A_i$ 是 X 上的模糊集,定义为

$$A(x)=\inf\{A_i(x_i):i\in I\},$$

这里 $x\in X$,其第 i 个坐标等于 x_i.

例7 设 $X=Y=[0,1]$,$A\in\mathscr{F}(X)$,$B\in\mathscr{F}(Y)$,

$$A(x)=x\qquad(x\in X),$$
$$B(y)=1-y\qquad(y\in Y),$$

则

$$A\times B\in\mathscr{F}(X\times Y),$$
$$(A\times B)(x,y)=x\wedge(1-y)\quad((x,y)\in X\times Y).$$

1.2 扩张原理与序同态

1.2.1 扩张原理

X 与 Y 的分明子集的概念已经分别扩充为 X 与 Y 的模糊子集概念,集 X 到集 Y 的映射 $f:X\to Y$ 就也应扩充为从 $\mathscr{F}(X)$ 到 $\mathscr{F}(Y)$ 的映射 $\bar{f}$.扎德提出的如下定义.

定义13(扩张原理) 设 $f:X\to Y$ 是映射,则得一映射 $\bar{f}:\mathscr{F}(X)\to\mathscr{F}(Y)$.$\forall A$

$\in \mathscr{F}(X)$，$\bar{f}(A) \in \mathscr{F}(Y)$，定义为

$$\bar{f}(A)(y) = \sup\{A(x):f(x) = y\} \quad (y \in Y), \tag{1-19}$$

这里约定[0,1]的空子集的上确界为零，$\bar{f}$有一逆映射$\bar{f}^{-1}:\mathscr{F}(Y) \to \mathscr{F}(X)$. $\forall B \in \mathscr{F}(Y)$，$\bar{f}^{-1}(B) \in \mathscr{F}(X)$，定义为

$$\bar{f}^{-1}(B)(x) = B(f(x)). \tag{1-20}$$

称$\bar{f}$为由f诱导出的**扎德型函数**.

为简便计，经常$\bar{f}$与$\bar{f}^{-1}$仍分别用f与f^{-1}表示.

定理2 设$f:\mathscr{F}(X) \to \mathscr{F}(Y)$是由$f: X \to Y$按扩张原理所得的扎德型函数，$f^{-1}:\mathscr{F}(Y) \to \mathscr{F}(X)$是其逆映射，则

1° $f(A) \leqslant B$，当且仅当$A \leqslant f^{-1}(B)$；

2° $f(\bigvee_{i \in I} A_i) = \bigvee_{i \in I} f(A_i)$，

$f^{-1}(B') = (f^{-1}(B))'$，

$f^{-1}(\bigvee_{i \in I} B_i) = \bigvee_{i \in I} f^{-1}(B_i)$，$f^{-1}(\bigwedge_{i \in I} B_i) = \bigwedge_{i \in I} f^{-1}(B_i)$；

3° $f^{-1}(B) = \bigvee_{i \in I}\{A \in \mathscr{F}(X):f(A) \leqslant B\}$，

$f(A) = \bigwedge_{i \in I}\{B \in \mathscr{F}(Y):A \leqslant f^{-1}(B)\}$，

$f^{-1}(B) = \bigwedge\{A \in \mathscr{F}(X):f(A') \leqslant B'\}$；

4° $A \leqslant f^{-1}f(A)$，$B \geqslant ff^{-1}(B)$；

5° $ff^{-1}f(A) = f(A)$，$f^{-1}ff^{-1}(B) = f^{-1}(B)$.

由定理2可看出，$f^{-1}:\mathscr{F}(Y) \to \mathscr{F}(X)$有较好的性质，即$f^{-1}$保并、保交，还保伪补，而$f:\mathscr{F}(X) \to \mathscr{F}(Y)$只保并，不保交，也不保伪补.

命题2 设$f:\mathscr{F}(X) \to \mathscr{F}(Y)$是由$f: X \to Y$诱导出的扎德型函数，$f^{-1}:\mathscr{F}(Y) \to \mathscr{F}(X)$是它的逆，则

1° $\forall \lambda \in [0,1]$，$f([\lambda]) = [\lambda]$，$f^{-1}([\lambda]) = [\lambda]$；

2° $f(\lambda A) = \lambda f(A)$；

3° $f^{-1}(\lambda B) = \lambda f^{-1}(B)$.

其中$[\lambda]$表示X上或Y上取常值λ的模糊集.

推论2

$$\begin{aligned} f(A) &= \bigvee_{r \in [0,1]} rf(A_r) = \bigvee_{r \in [0,1]} rf(A_{(r)}) \\ &= \bigvee_{r \in Q \cap [0,1]} rf(A_r) = \bigvee_{r \in Q \cap [0,1]} rf(A_{(r)}). \end{aligned} \tag{1-21}$$

$$\begin{aligned} f^{-1}(B) &= \bigvee_{r \in [0,1]} rf^{-1}(B_r) = \bigvee_{r \in [0,1]} rf^{-1}(B_{(r)}) \\ &= \bigvee_{r \in Q \cap [0,1]} rf^{-1}(B_r) = \bigvee_{r \in Q \cap [0,1]} rf^{-1}(B_{(r)}). \end{aligned} \tag{1-22}$$

1.2.2 序同态

定理2中列出了扎德型函数及其逆的一系列性质.但是反过来具有这些性质的映射并不一定是扎德型函数.前面已经说过$\mathscr{F}(X)$与$\mathscr{F}(Y)$作为单位区间的乘积都是具有逆序对合对应的完全分配格.可以把扎德型函数推广为这种格之间的映射.

定义 14 设 L 是具有逆序对合对应(伪补)的**分子格**,即完备的完全分配格,则称 L 为**模糊格**,简称 Fuzz.

定义 15 设 $f: L_1 \to L_2$ 是映射,这里 L_1 与 L_2 都是模糊格.如果

1° $f(\bigvee_{i \in I} a_i) = \bigvee_{i \in I} f(a_i)$,

2° $f^{-1}(b') = (f^{-1}(b))'$,

其中 $f^{-1}: L_2 \to L_1$ 的定义为

$$f^{-1}(b) = \vee \{a \in L_1 : f(a) \leqslant b\},$$

则称 f 为从 L_1 到 L_2 的**序同态**.

由定理 2 看出,扎德型函数是序同态.

定理 2 给出的扎德映射的那些性质都是序同态所具有的,即下面的定理 3 成立.

定理 3 设 $f: L_1 \to L_2$ 是序同态,这里 L_1 与 L_2 都是模糊格,则

1° $f(a) \leqslant b$,当且仅当 $a \leqslant f^{-1}(b)$;

2° f 保并,f^{-1} 保并,保交,保伪补;

3° $$f(a) = \wedge \{b \in L_2 : a \leqslant f^{-1}(b)\},$$
$$f^{-1}(b) = \wedge \{a \in L_1 : f(a') \leqslant b'\};$$

4° $a \leqslant f^{-1}f(a)$, $b \geqslant ff^{-1}(b)$;

5° $ff^{-1}f(a) = f(a)$,

$$f^{-1}ff^{-1}(b) = f^{-1}(b),$$

即
$$ff^{-1}f = f, \quad f^{-1}ff^{-1} = f^{-1}.$$

序同态不必为扎德型函数,关于序同态成为扎德型函数的充要条件可参看有关文献①.

1.3 L 模糊集

1.3.1 L 模糊集

X 上的模糊集 A 是一个函数 $A: X \to [0,1]$.设 $x \in X$,则 $A(x)$ 表示 x 对于 A 的隶属度,它是区间[0,1]中的一个实数,所以隶属度是清晰的且彼此之间是可以比较大小的.但在某些场合会涉及隶属度不很清晰和彼此不可比较的情形,这时可将区间[0,1]推广为模糊格而引入更广的模糊集概念.

定义 16 设 X 是非空集,L 是一个模糊格,则称从 X 到 L 的映射 $A: X \to L$ 为 X 上的 **L 模糊集**,简称为 LF 集.其全体记作 $\mathscr{F}_L(X)$ 或 L^X.

当 $L = [0,1]$ 时,L 模糊集就成为模糊集.

可以证明若干模糊格的乘积仍为模糊格,而 $\mathscr{F}_L(X)$ 是 $|X|$ 个 L 的乘积,所以有如下命题.

---

① Wang G J. Order homomorphisms on fuzzes. Fuzzy Sets and Systems, 1984(12): 281 - 288

命题 3 设 L 是模糊格，则 L^X 也是模糊格.

像模糊集一样，L 模糊集之间的并、交和伪补运算就是把它们看做映射时取上、下确界和伪补的运算.同时可对 L 模糊集定义截集与乘积.

定义 17 设 $A,B \in L^X, \forall i \in I, A_i \in L^X$.

1° 称 A 包含于 B，记作 $A \leqslant B$，若 $\forall x \in X, A(x) \leqslant B(x)$.这时也说 B 包含 A.

2° $(\bigvee_{i \in I} A_i)(x) = \sup_{i \in I} A_i(x), (\bigwedge_{i \in I} A_i)(x) = \inf_{i \in I} A_i(x), x \in X$.

3° $\lambda A = [\lambda] \wedge A$，其中 $[\lambda]$ 是在 X 上取常值 λ 的 L 模糊集，$\lambda \in L$.

4° $A'(x) = (A(x))', x \in X$. A' 叫 A 的伪补.

5° 称 $A_r = \{x \in X: A(x) \geqslant r\}$ 为 A 的 r 截集，$r \in L$.

6° $A \times B \in \mathscr{F}_L(X \times X), \forall (x,y) \in X \times X, (A \times B)(x,y) = A(x) \wedge B(y)$.

对 L 模糊集而言，有下述的定理、定义.

定理 4 设 $A \in L^X$，则

$$A = \bigvee \{rA_r : r \in L\}. \tag{1-23}$$

设 $r \in L$，一般 $r \neq \sup\{s \in L: s < r\}$，所以(1-23)式中的 r 截集 A_r 不能改为 r 强截集.

设 $f: X \to Y$ 是映射，则 f 可像定义 12 一样诱导出两个映射来.

定义 18 设 $f: X \to Y$ 是映射，则 f 诱导一个映射 $f: L^X \to L^Y$. $\forall A \in L^X, f(A) \in L^Y$，定义为

$$f(A)(y) = \sup\{A(x): f(x) = y\}, y \in Y.$$

称此映射为 **L 型扎德映射**.其逆映射 $f^{-1}: L^Y \to L^X$，定义为

$$f^{-1}(B)(x) = B(f(x)) \quad (x \in X),$$

其中 $B \in L^Y$.

可以证明 L 型扎德函数也是序同态，从而定理 2 中的那些性质对 L^X 中的 A 与 A_i 以及 L^Y 中的 B 与 B_i 也成立.

1.3.2 区间值模糊集

所谓区间值模糊集是一种特殊的 L 模糊集.下面先介绍区间值模糊集的原始定义及其运算，然后论证它实际上是一种 L 模糊集.

定义 19 设 $l = \{[a,b]\}: 0 \leqslant a \leqslant b \leqslant 1\}$.规定

1° $[a,b] \leqslant [c,d]$，当且仅当 $a \leqslant c$ 且 $b \leqslant d$；

2° $\bigvee_{i \in I} [a_i, b_i] = [\bigvee_{i \in I} a_i, \bigvee_{i \in I} b_i]$,

$\bigwedge_{i \in I} [a_i, b_i] = [\bigwedge_{i \in I} a_i, \bigwedge_{i \in I} b_i]$；

3° $[a,b]' = [b', a']$.

称 l 为**区间值集**.

定义 20 设 X 是非空集，则从 X 到 l 的映射 $A: X \to l$ 称为**区间值模糊集**.区间值模糊集之间的序以及并、交、伪补运算均按定义 17 的形式根据定义 19 而逐点定义.

可以证明，区间值模糊集就是对于适当选取的模糊格 L 而言的 L 模糊集.事实

上,$[0,1]^2$ 是一个模糊格,令

$$L=\{(x,y)\in[0,1]^2:x\leqslant y\},$$
$$(x,y)'=(y',x').\tag{1-24}$$

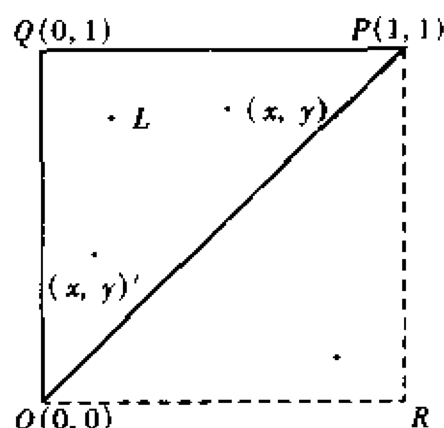


图 1-6

即规定 L 中的序与 $[0,1]^2$ 中的序相同,但规定 $(x,y)'=(y',x')$,则 L 成为一个模糊格,其中的序、并、交以及伪补运算均与定义 19 相吻合.因此,令 L 中的点 (x,y) 与 I 中的闭区间 $[x,y]$ 相对应,则得一 L 与 I 之间的同构.把同构的对象不加区别时,有如下定理.

定理 5 区间值模糊集就是 L 模糊集,这里的 L 由(1-24)式确定.

L 的图像如图 1-6 所示.L 的最大元为 $P(1,1)$、最小元为原点 $O(0,0)$.点 (x,y) 及其伪补 $(x,y)'=(y',x')$ 关于正方形 $ORPQ$ 的另一条对角线 QR 对称.

1.3.3 直觉主义模糊集

1984 年保加利亚的阿坦纳索夫(Atanassov)引入了所谓直觉主义模糊集(intuitionistic fuzzy sets)的概念如下:

定义 21 设 X 是非空集,称

$$A^*=\{\langle x,\mu_A(x),\nu_A(x)\rangle:x\in X\}\tag{1-25}$$

为 X 上的**直觉主义模糊集**,其中 $\mu_A:X\to[0,1]$ 与 $\nu_A:X\to[0,1]$ 分别表示 x 对于 A 的隶属度和非隶属度,满足条件:

$$0\leqslant\mu_A(x)+\nu_A(x)\leqslant 1\quad(x\in X).\tag{1-26}$$

可以证明,直觉主义模糊集就是对于适当选取的模糊格 L 而言的 L 模糊集.事实上,$[0,1]$ 与 $[0,1]^{\mathrm{OP}}$ 都是模糊格,这里 $[0,1]^{\mathrm{OP}}$ 表示把 $[0,1]$ 中的序颠倒过来所得的模糊格.那么由模糊格的乘积为模糊格可知,$[0,1]\times[0,1]^{\mathrm{OP}}$ 是一个模糊格.令

$$L=\{(x,y)\in[0,1]\times[0,1]^{\mathrm{OP}}:x+y\leqslant 1\}.\tag{1-27}$$

规定 L 中的序与 $[0,1]\times[0,1]^{\mathrm{OP}}$ 中的序相同,即

$$(x,y)\leqslant(u,v)\quad(\text{当且仅当 }x\leqslant u\text{ 且 }v\leqslant y).\tag{1-28}$$

但规定

$$(x,y)'=(y,x).\tag{1-29}$$

则 L 成为一个模糊格,其中的序、并、交以及伪补运算均与阿坦纳索夫给出的相吻合.

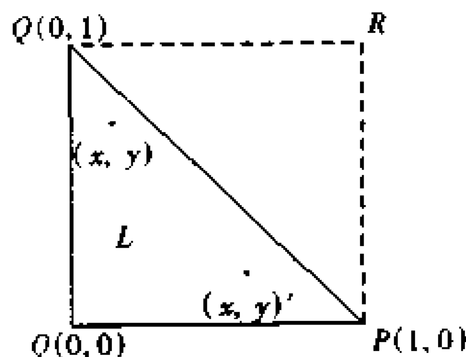


图 1-7

注意(1-25)式无非是把自变量 x 与 x 所对应的函数值 $(\mu_A(x),\nu_A(x))$ 写在了一起,并另用 A^* 表示这个函数的一种复杂记法而已.简单地说,(1-25)式确定了一个从 X 到 L 的映射 A,这里 L 由(1-27)式确定,所以有如下定理.

定理 6 直觉主义模糊集就是 L 模糊集,这里 L 由

(1-27) 式与(1-29) 式确定.

L 的图像如图 1-7 所示. L 的最大元为 $P(1,0)$,最小元为 $Q(0,1)$. 点(x,y)及其伪补$(x,y)'=(y,x)$关于正方形 $OPRQ$ 的另一条对角线 OR 对称.

1.3.4 Ⅱ 型模糊集

定义 22 设 X 是非空集,则称从 X 到 $\mathscr{F}([0,1])$ 的映射 $A: X \to \mathscr{F}([0,1])$ 为 X 上的 Ⅱ **型模糊集**.

显然 $L=\mathscr{F}([0,1])$ 是一个模糊格,所以有如下定理.

定理 7 **Ⅱ 型模糊集**就是 L 模糊集,这里 $L=\mathscr{F}([0,1])$.

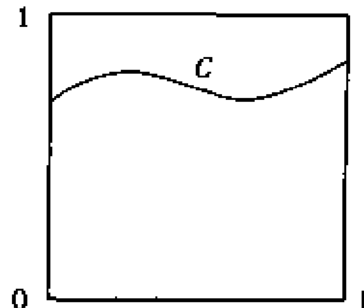


图 1-8

L 的图像如图 1-8 所示. 这时 L 的一个点由图 1-8 所示的一条曲线 C 表示.

1.4 论域上带有结构的模糊集

1.4.1 论域上的结构

虽然在模糊集的定义中并不要求在论域 X 上带有结构,但在实用上往往遇到在 X 上带有序结构或代数结构或拓扑结构的情形.

定义23 设 X 为向量空间, A 是 X 上的模糊集,若对于 X 中任二点 x 与 y,以及 $[0,1]$ 中的任一实数 t,

$$A(tx+(1-t)y) \geqslant A(x) \wedge A(y) \tag{1-30}$$

恒成立,则称 A 为 X 上的**凸模糊集**.

特别当 X 是实直线或区间时,由(1-30)式表明, A 在任一区间中的值不小于它在区间两端的值的较小者,由此可知凡单调函数都是凸的(如图 1-9 所示,后两图为单调函数).

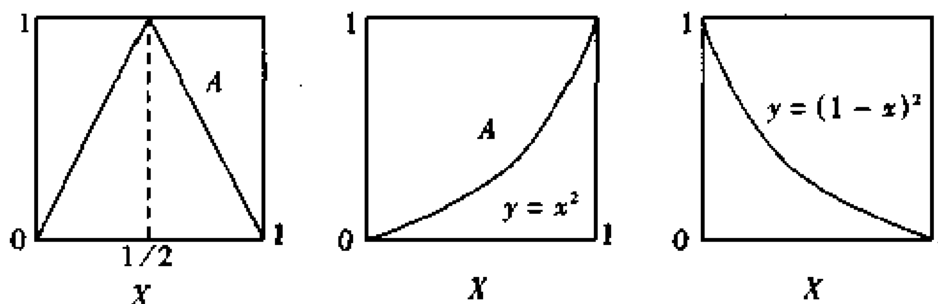


图 1-9

定义 24 设$(X,\mathscr{U})$是拓扑空间, A 是 X 上的模糊集. 若对于每个 $r\in[0,1]$, A 的 r 强截集 $A_{(r)}$ 都是$(X,\mathscr{U})$中的开集,则称 A 为$(X,\mathscr{U})$上的**下半连续函数**.

容易验证:

1° X 与 $\varnothing$ 是下半连续函数.

2° 若 A 与 B 是下半连续函数,则 $A \wedge B$ 也是下半连续函数.

3° 若 $\forall i \in I, A_i$ 是下半连续函数,则 $\bigvee_{i \in I} A_i$ 也是下半连续函数.

定义 25 设 X 是半群,A 是 X 的模糊子集,如果对于 X 的任二元 x 与 y,恒有

$$A(x \cdot y) \geqslant A(x) \wedge A(y) \tag{1-31}$$

成立,则称 A 为 X 的**模糊子半群**.

1.4.2 模糊数

模糊数是实数概念的推广.模糊数的定义不很统一,下面采取与文献[1]中的定义等价的定义.

定义 26 设 A 是 **R** 上的模糊集,如果

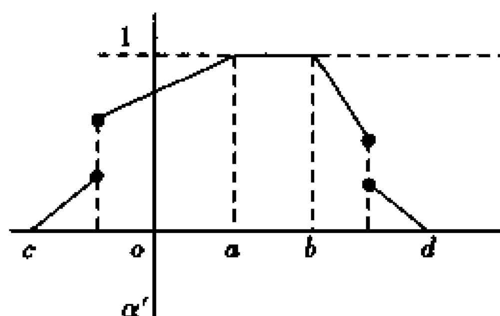


图 1-10

1° A 是**正规模糊集**,即 A 于论域中某点处的值等于 1,

2° A 是 **R** 上的凸模糊集,

3° $\forall r \in [0,1]$,A 的 r 截集 A_r 是有界闭区间,

4° $\mathrm{supp}A$ 为有界区间,

则称 A 为**模糊数**,简称为 **F 数**.

注 2 (1) 也有的著作称 **R** 上满足以上两个条件 1° 与 2° 或 1° 与 3° 的模糊集为模糊数,如文献[2]、[3]等.

(2) 定义 26 中所说的闭区间可以是蜕化的,即可以是左右端点重合的由一个点组成的闭区间.所以对于任一实数 α,α 的特征函数是模糊数.

(3) 可以证明模糊数的构造如图 1-10 所示.设此模糊数为 A,则 A 在闭区间 $[a,b]$ 上的值等于 1(可能 $a = b$),在区间 $[c,a]$ 上递增(不减)且右连续,在区间 $[b,d]$ 上递减(不增)且左连续.A 的承集 $\mathrm{supp}A$ 为区间(开或闭或半开半闭),此图中 $\mathrm{supp}A = (c,d)$.

定义 27 设 $*$ 是 **R** 上的二元运算,则 $*$ 可扩张成 **R** 上模糊集的运算.设 $A, B \in \mathscr{F}(\mathbf{R})$,则 $A * B \in \mathscr{F}(\mathbf{R})$,定义为

$$(A * B)(z) = \bigvee_{x * y = z} [A(x) \wedge B(y)]. \tag{1-32}$$

特别当 A, B 为模糊数时,可按(1-32)式定义它们的和与积:

$$(A + B)(z) = \bigvee_{x + y = z} [A(x) \wedge B(y)], \tag{1-33}$$

$$(A \cdot B)(z) = \bigvee_{xy = z} [A(x) \wedge B(y)]. \tag{1-34}$$

又设 k 为任一实数,则规定当 $k \neq 0$ 时,$(kA)(z) = A(z/k)$;当 $k = 0$ 时,$kA = 0$.

注 3 (1) 对于模糊数而言,还可根据(1-32)式引入减法与除法运算,但它们并不是加法与乘法的逆运算,即 $(A - B) + B$ 不必等于 A,$(A \div B) \times B$ 也未必等于 A(参看文献[3]).

(2) 可以证明:当 A 与 B 为模糊数时,$A + B, A \cdot B$ 以及 kA 也为模糊数(参看文

献[3]).

(3) $k(A+B)=kA+kB$, $k_1(k_2A)=(k_1k_2)A$, 且当 $k_1\geqslant 0$, $k_2\geqslant 0$ 时, $(k_1+k_2)A=k_1A+k_2A$(参看文献[1]).

模糊数的运算和它的截集闭区间的运算紧密相关.闭区间的运算有如下定义、定理.

定义 28 设 $[a,b]$, $[c,d]$ 是闭区间, k 是实数,则

1° $[a,b]+[c,d]=[a+c,b+d]$.

2° $[a,b]\cdot[c,d]=[\alpha,\beta]$,

其中

$$\alpha=\min\{ac,ad,bc,bd\},$$
$$\beta=\max\{ac,ad,bc,bd\}.$$

3° 当 $k\geqslant 0$ 时, $k[a,b]=[ka,kb]$; 当 $k<0$ 时, $k[a,b]=[kb,ka]$.

定理 8 设 A, B 为模糊数, k 为实数,则 $A+B$, $A\cdot B$ 与 kA 为模糊数,且 $\forall r\in[0,1]$.

1° $(A+B)_r=A_r+B_r$,

2° $(A\cdot B)_r=A_r\cdot B_r$,

3° $(kA)_r=kA_r$.

例 8 设模糊数 A 的表达式为

$$A(x)=\begin{cases}x-1 & (1\leqslant x\leqslant 2),\\ 3-x & (2<x\leqslant 3),\\ 0 & (x<1 \text{ 或 } x>3).\end{cases} \tag{1-35}$$

则

$$(A+A)(z)=\bigvee_{x+y=z}[A(x)\wedge A(y)]=\bigvee_{x}[A(x)\wedge A(z-x)].$$

按上式求 $A+A$ 的表达式较繁.方便的办法是利用定理 8.由 A 的隶属函数(1-35)式可知, $\forall r\in(0,1]$,

$$A_r=[1+r,3-r].$$

所以

$$(A+A)_r=A_r+A_r=[2+2r,6-2r].$$

令 $z=2+2r$, 得 $r=\frac{z}{2}-1$. 令 $6-2r=z$ 得 $r=3-\frac{z}{2}$. 注意 $(A+A)_{(0)}=(2,6)$, $(A+A)_1=[4,4]$, 便得出 $A+A$ 的隶属数为

$$(A+A)(z)=\begin{cases}\frac{z}{2}-1 & (2\leqslant z\leqslant 4),\\ 3-\frac{z}{2} & (4<z\leqslant 6),\\ 0 & (z<2 \text{ 或 } z>6).\end{cases} \tag{1-36}$$

顺便指出,如果利用定理8求 $2A$,则可得出(1-36)式同样的表达式,所以 $A+A=2A$.

2　若干理论分支

2.1　模糊测度与模糊积分

定义1　设 X 是非空集，$\mathscr{A}\subseteq\mathscr{P}(X)$. 若 $X\in\mathscr{A}$，且 $\mathscr{A}$ 对补运算和可数并运算封闭，则称 $\mathscr{A}$ 为 X 上的 σ 代数.

显见这时 $\varnothing\in\mathscr{A}$，且由对偶律可知，$\mathscr{A}$ 对可数交运算封闭. 特别是，当 A,B 属于 $\mathscr{A}$ 时，$A\cup B$，$A\cap B$ 与 $A-B$ 均属于 $\mathscr{A}$，这时称 $(X,\mathscr{A})$ 为可测空间.

2.1.1　模糊测度与可测函数

关于模糊测度与模糊积分，存在若干不同的理论，本篇介绍的是被普遍接受的一种。

定义2　设 $(X,\mathscr{A})$ 为可测空间，$\mu:\mathscr{A}\to[0,1]$ 是映射. 如果 $\mu(\varnothing)=0$，$\mu(X)=1$，当 $A,B\in\mathscr{A}$，且 $A\subseteq B$ 时，

$$\mu(A)\leqslant\mu(B),$$

以及当 $A_1\subseteq A_2\subseteq\cdots$ 时，

$$\lim_{n\to\infty}\mu(A_n)=\mu\left(\bigcup_{n=1}^{\infty}A_n\right),$$

当 $A_1\supseteq A_2\supseteq\cdots$ 时，

$$\lim_{n\to\infty}\mu(A_n)=\mu\left(\bigcap_{n=1}^{\infty}A_n\right),$$

则称 μ 为 $(X,\mathscr{A})$ 上的**模糊测度**，简称为 F 测度，称 $(X,\mathscr{A},\mu)$ 为**模糊测度空间**，简称为 F 测度空间.

注意，在模糊测度的定义中并没有任何模糊集出现，这里只是把通常测度定义中的可加性弱化为单调性而已.

定义3　设 μ 是 $(X,\mathscr{A})$ 上的模糊测度，$-1<\lambda<\infty$. 若当 $A,B\in\mathscr{A}$，且 $A\cap B=\varnothing$ 时，

$$\mu(A\cup B)=\mu(A)+\mu(B)+\lambda\mu(A)\mu(B),\tag{2-1}$$

则称 μ 为**λ 可加测度**. 这时也常把 μ 写为 g_λ. (2-1) 式也可推广为 A 与 B 交非空的情形，这时有

$$\mu(A\cup B)=\frac{\mu(A)+\mu(B)+\lambda\mu(A)\mu(B)-\mu(A\cap B)}{1+\lambda\mu(A\cap B)}.\tag{2-2}$$

注意，由 $\mu(\varnothing)=0$ 可知，(2-1) 式是 (2-2) 式中 $A\cap B=\varnothing$ 时的特殊情形.

易证，对 λ 可加的模糊测度 μ 而言，

$$\mu(X-A)=\frac{1-\mu(A)}{1+\lambda\mu(A)}.\tag{2-3}$$

定义4　$(X,\mathscr{A})$ 上的模糊测度 μ 称为**自连续**的，若 $\forall A\in\mathscr{A}$，以及对于一切满足 $\mu(B_n)\to0$ 的序列 $\{B_n\}\subseteq\mathscr{A}$，恒有

$$\mu(A\cup B_n)\to\mu(A),\quad\mu(A-B_n)\to\mu(A).\tag{2-4}$$

设$(X,\mathscr{A})$是可测空间，$f: X \to [0,1]$是映射．如果$\forall c \in [0,1]$，

$$\{x \in X: f(x) > c\} \in \mathscr{A},$$

则称f为$(X,\mathscr{A})$上的可测函数．

定义 5　设f与$f_n(n=1,2,\cdots)$都是$(X,\mathscr{A})$上的可测函数，$A \in \mathscr{A}$，如果存在$E \in \mathscr{A}$使$\mu(E)=0$且$\forall x \in A-E$，$f_n(x) \to f(x)$，则称$\{f_n\}$在A上**几乎处处收敛**于f.

如果$\forall \varepsilon > 0$，$\mu\{x \in A: |f_n(x)-f(x)| > \varepsilon\} \to 0$，则称$\{f_n\}$在$A$上**依测度$\mu$收敛**于$f$.

定理 1　设f与$f_n(n=1,2,\cdots)$都是$(X,\mathscr{A})$上的可测函数，μ是$(X,\mathscr{A})$上的模糊测度，$A \in \mathscr{A}$，则

1° 当$\{f_n\}$在A上几乎处处收敛于f时，$\{f_n\}$在A上依测度μ收敛于f；

2° 设μ是自连续的，则当$\{f_n\}$在A上依测度μ收敛于f时，$\{f_n\}$有子列在A上几乎处处收敛于f；

3° 设μ是自连续的，且$\{f_n\}$在A上几乎处处收敛于f，则$\forall \varepsilon > 0$，存在$E \in \mathscr{A}$，$\mu(E) < \varepsilon$，$\{f_n\}$在$A-E$上一致收敛于f.

2.1.2　模糊积分

定义 6　设$(X,\mathscr{A},\mu)$是模糊测度空间，$f: X \to [0,1]$是可测函数，$A \in \mathscr{A}$，令

$$\int_A f \mathrm{d}\mu = \sup_{E \in \mathscr{A}} \left(\inf_{x \in E} f(x) \wedge \mu(A \cap E) \right), \tag{2-5}$$

称$\int_A f(x)\mathrm{d}\mu$为f在A上的**模糊积分**，简称为**F 积分**.

模糊积分的几何解释如图 2-1 所示．由图可见，如果在$\mathscr{A}$中选取位于f的图像顶峰下面的很小的可测集E，则可使$\inf\limits_{x \in E} f(x)$很高(大)，但这时$\mu(A \cap E)$却很小，从而括号内$(\inf\limits_{x \in E} f(x) \wedge \mu(A \cap E))$的值很小，而$f$在$A$上的模糊积分正是一切可能的这种括号内的值的上确界.

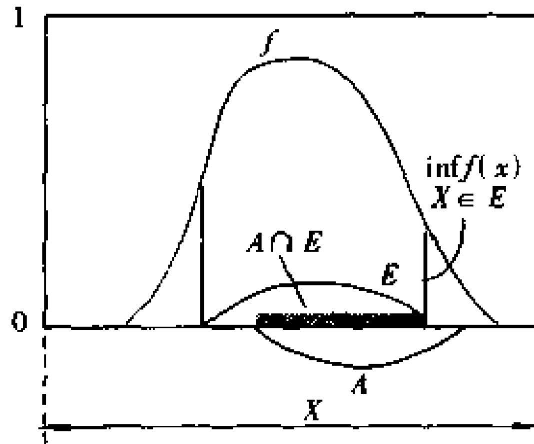


图 2-1

定理 2　设$(X,\mathscr{A},\mu)$是模糊测度空间，$f: X \to [0,1]$是可测函数，$A \in \mathscr{A}$，则f在A上的模糊积分可表示为

$$\int_A f \mathrm{d}\mu = \sup_{\alpha \in [0,1]} [\alpha \wedge \mu(A \cap f_\alpha)], \tag{2-6}$$

其中f_α表示f作为模糊集看待时的α截集.

(2-6) 式的几何意义如图 2-2 所示．由图可见，α越大，则f_α越小，从而括号内$(\alpha \wedge \mu(A \cap f_\alpha))$的值也越小，而$f$在$A$上的模糊积分正是一切可能的这种括号内的值的上确界．特别当积分区域A是整个空间X时，f的模糊积分值等于内接于f的

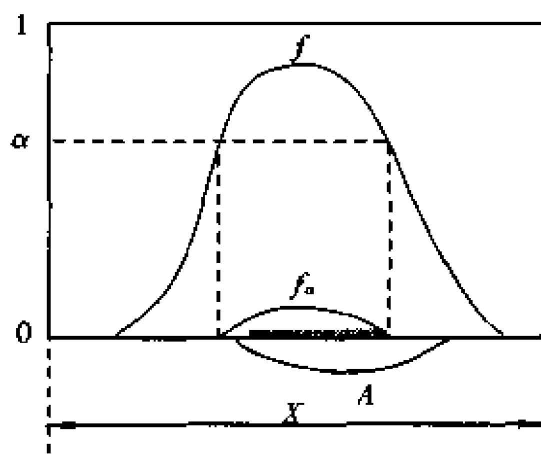


图 2-2

图像的正方形的边长.

注 1　如果把(2-6)式中括号里的交运算 $\wedge$ 改为乘积运算,则得所谓的 N 积分概念. 在积分区域 A 等于 X 时, f 的 N 积分的值等于 f 的图像下方的最大矩形的面积.

定理 3　模糊积分具有以下性质:

1° 两种单调性,即

当 $f \leqslant g$ 时, $\int_A f\mathrm{d}\mu \leqslant \int_A g\mathrm{d}\mu$;

当 $A \subseteq B$ 时, $\int_A f\mathrm{d}\mu \leqslant \int_B f\mathrm{d}\mu$.

2°　$\int_A c\mathrm{d}\mu = c \wedge \mu(A)$　$(c \in [0,1])$;

$$\int_X (\lambda \wedge f)\mathrm{d}\mu = \lambda \wedge \int_X f\mathrm{d}\mu.$$

3° 设可测函数列 $\{f_n\}$ 单调($\{f_n\}$ 为递增或递减函数列)收敛于 f,则

$$\lim_{n\to\infty}\int_A f_n\mathrm{d}\mu = \int_A f\mathrm{d}\mu.$$

4° 设模糊测度 μ 是次可加的,即条件

$$\mu(A \cup B) \leqslant \mu(A) + \mu(B)$$

成立,且可测函数列 $\{f_n\}$ 收敛于 f,则

$$\lim_{n\to\infty}\int_A f_n\mathrm{d}\mu = \int_A f\mathrm{d}\mu.$$

关于 L 模糊测度与 L 模糊积分的进一步知识可参看文献[4].

2.2　L 模糊代数

2.2.1　L 模糊关系

定义 7　设 L 是完备格, $T: L^2 \to L$ 是 L 上的二元运算,如果 T 是交换的,结合的和单调的,即当 $x,y,z \in L$ 时, $T(x,y) = T(y,x)$, $T(T(x,y),z) = T(x,T(y,z))$ 和当 $x \leqslant y$ 时, $T(x,z) \leqslant T(y,z)$ 成立,则分别当 $T(1,x) = x$ 或 $T(0,x) = x$ 成立时,称 T 为三角模或余三角模. 余三角模也常被称为 S 模,并以 S 记之. 而三角模也简称为 T 模.

例 1　设 $L = [0,1]$. 定义

$$T_0(a,b) = \begin{cases} a & (b = 1), \\ b & (a = 1), \\ 0 & (\text{其他}); \end{cases}$$

$$T_0(a,b) = a \wedge b;$$

$$T_1(a,b) = ab;$$
$$T_2(a,b) = \frac{ab}{1+(1-a)(1-b)};$$
$$T_\infty(a,b) = (a+b-1) \vee 0.$$

以上5个函数是 T 模,

$$S'_0(a,b) = \begin{cases} a & (b=0), \\ b & (a=0), \\ 1 & (\text{其他}). \end{cases}$$

定义

$$S_0(a,b) = a \vee b,$$
$$S_1(a,b) = a+b-ab,$$
$$S_2(a,b) = \frac{a+b}{1+ab},$$
$$S_\infty(a,b) = (a+b) \wedge 1.$$

以上5个函数是 S 模,且

$$T'_0 \leqslant T_\infty \leqslant T_2 \leqslant T_1 \leqslant T_0 \leqslant S_0 \leqslant S_1 \leqslant S_2 \leqslant S_\infty \leqslant S'_0.$$

可以证明,T_0 与 T'_0 分别是最大与最小的 T 模,S_0 与 S'_0 分别是最小与最大的 S 模.

定义 8　设 X 与 Y 是非空集,L 是模糊格,$R \in L^{X\times Y}$,即 R 是 $X\times Y$ 上的 L 模糊集,则称 R 为从 X 到 Y 的 **L 模糊关系**,简称为 **LF 关系**.这时设

$$R^{-1}(y,x) = R(x,y) \qquad ((y,x) \in Y\times X),$$

则 R^{-1} 是从 Y 到 X 的 L 模糊关系,称为 R 的**逆关系**.

设 $Q \in L^{X\times Y}$, $R \in L^{Y\times Z}$.定义 $Q\circ R \in L^{X\times X}$ 如下:

$$(Q\circ R)(x,z) = \bigvee_{y\in Y} Q(x,y) \wedge R(y,z),(x,z) \in X\times Z. \tag{2-7}$$

称 $Q\circ R$ 为 Q 与 R 的复合关系,简称复合.

(2-7)式中的 $Q(x,y) \wedge R(y,z)$ 也可用更广泛的 T 模表达为 $T(Q(x,y), R(y,z))$,或简写为 $Q(x,y)TR(y,z)$.这样就得到更一般的 T 复合关系,并把这种复合运算记为 O_T.为简明计,本篇不讨论这种一般的复合关系,但读者不难在以下的讨论中把两个 L 模糊集的交运算换成 T 模运算(或再加适当的条件),而得出更一般的结论来.

命题 1　设 $Q \in L^{X\times Y}$, $R, R_i \in L^{Y\times Z}(i \in I)$, $S \in L^{Z\times W}$,则

1° $(Q\circ R)^{-1} = R^{-1}\circ Q^{-1}$,

2° $(Q\circ R)\circ S = Q\circ(R\circ S)$,

3° $Q_\lambda \circ R_\lambda \subseteq (Q\circ R)_\lambda, \lambda \in L$,

4° $Q\circ(\bigvee_{i\in I} R_i) = \bigvee_{i\in I}(Q\circ R_i)$,

5° $(\bigvee_{i\in I} R_i)\circ S = \bigvee_{i\in I}(R_i\circ S)$.

考虑形如

$$Q\circ R = S, \quad Q \in L^{X\times Y}, \quad R \in L^{Y\times Z}, \quad S \in L^{X\times Z} \tag{2-8}$$

的等式.如果 Q 与 S 是已知的,在 $L^{Y\times Z}$ 中求 R 使(2-8)式成立,则称(2-8)式为关于

R 的 L 模糊关系方程.满足(2-8)式的 R 叫(2-8)式的解.自然 R 也可处于复合运算中第一个因子的位置,但由 $(R\circ Q)^{-1}=Q^{-1}\circ R^{-1}$ 可知,这种情况可以转化为形如(2-8)式的关系方程.以下用 $\mathscr{R}$ 记关系方程(2-8)的全部解之集.

设 $\alpha:[0,1]^2\to[0,1]$ 是**哥德尔(K.Gödel)蕴涵算子**,即

$$\alpha(a,b)\cdot=\begin{cases}1 & (a\leqslant b),\quad (a,b)\in[0,1]^2,\\ b & (a>b).\end{cases}\tag{2-9}$$

这时也将 $\alpha(a,b)$ 记为 $a\alpha b$.

定义 9　设 $Q\in L^{X\times Y}$, $R\in L^{Y\times Z}$.定义 $Q\alpha R\in L^{X\times Z}$ 如下:

$$Q\alpha R\in\bigwedge_{y\in Y}Q(x,y)\alpha R(y,z).\tag{2-10}$$

称 $Q\alpha R$ 为 Q 与 R 的 α 复合.

命题 2　设 $Q\in L^{X\times Y}$, $R,R_i\in L^{X\times Z}(i\in I)$, $S,S_j\in L^{Z\times W}(j\in J)$,则

1° $(Q\alpha R)\alpha S=Q\alpha(R\alpha S)$,

2° $(\bigvee_{i\in I}R_i)\circ S=\bigwedge_{i\in I}(R_i\alpha S)$,

3° $R\alpha(\bigwedge_{j\in J}(S_j)=\bigwedge_{j\in J}(R\alpha S_j)$.

定理 4　若 L 模糊关系方程(2-8)式的解集 $\mathscr{R}$ 非空,则

1° $R^*=Q^{-1}\alpha S\in\mathscr{R}$,且 R^* 是(2-8)式的最大解.

2° 设 $R_1,R_2\in\mathscr{R}$,且 $R_1\leqslant R\leqslant R_2$,则 $R\in\mathscr{R}$.

3° 设 $\forall i\in I$, $R_i\in\mathscr{R}$,则 $\bigvee_{i\in I}R_i\in\mathscr{R}$.

又,当 Q 与 S 给定时,以 $\mathscr{R}^{\#}$ 记

$$Q\alpha R=S\tag{2-11}$$

的全部解之集,则当 $\mathscr{R}^{\#}$ 非空时,有

4° $R_*=Q^{-1}\circ S$ 是(2-11)式的最小解.

5° 设 $R_1,R_2\in\mathscr{R}^{\#}$,且 $R_1\leqslant R\leqslant R_2$,则 $R\in\mathscr{R}^{\#}$.

6° 设 $\forall i\in I$, $R_i\in\mathscr{R}^{\#}$,则 $\bigwedge_{i\in I}R_i\in\mathscr{R}^{\#}$.

以下考虑 X 上的 L 模糊关系,即从 X 到其自身的 L 模糊关系.

定义 10　设 $R\in L^{X\times X}$,则分别当 $R^{-1}=R$, $E_X\leqslant R$ 和 $R\circ R\leqslant R$ 时,称 R 是对称的、自反的和传递的,这里 $E_X\in L^{X\times X}$,定义为

$$E_X(x,y)=\begin{cases}1 & (\text{当 } x=y,x,y\in X),\\ 0 & (\text{当 } x\neq y,x,y\in X).\end{cases}$$

$R\vee R^{-1}$ 与 $R\vee E_X$ 显然分别是包含 R 的最小的对称和自反的关系,分别称为 R 的对称闭包与自反闭包.以下以 $\mathscr{R}$ 记 X 上的全体 L 模糊关系之集.又称 X 上的对称、自反且传递的 L 模糊关系为 X 上的等价 L 模糊关系.

由命题1可知,复合运算"∘"是结合的,所以

$$(R\circ R)\circ R=R\circ(R\circ R),$$

以 R^3 记此复合的结果.一般地,以 R^n 记 n 个 R 复合的结果,它是与括号的位置无关的.显然

$$R^m\circ R^n=R^{m+n}.$$

又,当 R 对称时,R^n 也对称,当 R 自反时,R^n 也自反.

关于传递性,情况稍稍复杂一些,有下面的定理.

定理 5 设 R 是 X 上的 L 模糊关系,则

1° R 是传递的,当且仅当 $\forall \lambda \in L, R_\lambda$ 是 X 上传递的分明二元关系.

2° 若 R 是传递的,则 $\forall n \in N, R^n$ 也是传递的.

3° $\mathscr{R}$ 中若干传递关系之交是传递的.又 $1_{X\times X}$ 是 $\mathscr{R}$ 中最大的传递关系,从而 $\mathscr{R}$ 中有一包含 R 的最小传递关系,称为 R 的**传递闭包**.

4° 设 R 是自反的,则 R 传递的充要条件为

$$R^2 = R.$$

5° R 的传递闭包等于 $\bigvee\limits_{n=1}^{\infty} R^n$.

6° R 是 X 上的 L 模糊等价关系,当且仅当 $\forall \lambda \in L, R_\lambda$ 是 X 上的分明等价关系.

7° 存在包含 R 的最小等价 L 模糊关系,称为 R 的**等价闭包**,其结构是 $\bigvee\limits_{n=1}^{\infty}(R^*)^n$,这里,

$$R^* = (R \vee E_X) \vee (R \vee E_X)^{-1}.$$

定义 11 设 R 是 X 上的等价 L 模糊关系,$\forall x \in X$,定义 X 上的 L 模糊集 R_x 如下:

$$R_x(y) = R(x,y) \quad (y \in X).$$

称 R_x 为 x **陪集**;称 $\{R_x : x \in X\}$ 为 X 关于 R 的**商集**,记作 $\frac{X}{R}$.

R_x 实际上是 R 在 (x,X) 上的限制.因为 R 是对称的,也可将 R_x 定义为 R 在 (X,x) 上的限制.易证,$R_x = R_y$,当且仅当 $R(x,y) = 1$;$R_x \wedge R_y = 0$,当且仅当 $R(x,y) = 0$.

注意:当 X 与 Y 都是有限集时,从 X 到 Y 的 L 模糊关系 $R \in L^{X\times Y}$ 可写作一个矩阵 $[r_{ij}]$,这里

$$r_{ij} = R(x_i, y_j) \quad (i = 1,2,\cdots,n, \quad j = 1,2,\cdots,m).$$

称 $[r_{ij}]$ 为 n 行 m 列的 **L 模糊矩阵**,或 $n \times m$ L 模糊矩阵.可见 L 模糊矩阵是 L 模糊关系的特殊情况.

2.2.2 L 模糊子代数

设 A 是 Ω 型的泛代数,$\forall n \in N, \Omega(n)$ 是 A 上的 n 元运算之集(可能是空集,$n = 1,2,\cdots$).又 $\Omega(0)$ 表示 A 中特定常元之集,$\Omega = \bigcup\limits_{k=0}^{\infty}\Omega(k)$.

定义 12 设 A 是 Ω 型泛代数 $\mu_i \in L^A (i = 1,2,\cdots,n)$,$\omega \in \Omega(n)$,定义 $\nu \in L^A$ 如下:

$$\nu(x) = \bigvee\left\{\bigwedge_{i=1}^{n}\mu(x_i) : \omega(x_1,x_2,\cdots,x_n) = x; x_1,x_2,\cdots,x_n \in A\right\} \quad (x \in A). \tag{2-12}$$

称 ν 为 $\mu_1,\mu_2,\cdots,\mu_n$ 的 **ω 乘积**.设 $\mu \in L^A$,如果

1° 当 $\omega \in \Omega(0)$ 时，$\mu(\omega) = 1$,

2° 当 $\omega \in \Omega(n)$，且 $x_1, x_2, \cdots, x_n \in A$ 时，

$$\mu(\omega(x_1, x_2, \cdots, x_n)) \geqslant \bigwedge_{i=1}^{n} \mu(x_i),$$

则称 μ 为 A 的 **L 模糊子代数**，简称为 A 的 LF 子代数.其全体记作 $L(A)$.

定理 6 $L(A)$ 按 L 模糊集的包含序构成一完备格，且其中求下确界的运算就是 L^A 中的普通交运算.又 1_A 是$L(A)$ 中的最大元，所以 $\forall \mu \in L^A$, $L(A)$ 中有一包含 μ 的最小 L 模糊子代数，称为由 μ 生成的 L 模糊子代数，记作$\langle\mu\rangle$.又 $\mu_\lambda \in L(A)$ 的充要条件为 $\forall \lambda \in L$, μ_λ 是 A 的分明子代数.

定义 13 设 A, B 都是 Ω 型的泛代数.

1° 设 $f: A \to B$ 是满同态，$\mu \in L(A)$, $\nu \in L(B)$.如果 $f(\mu) \leqslant \nu$，则称 f 为从 μ 到 ν 的**弱同态**，或称 μ 弱同态于 ν，记作 $\mu \overset{f}{\sim} \nu$.如果 $f(\mu) = \nu$，则称 f 为从 μ 到 ν 的**同态**，或称 μ 同态于 ν，记作 $\mu \overset{f}{\approx} \nu$.

2° 设 $f: A \to B$ 是同构，$\mu \in L(A)$, $\nu \in L(B)$，则分别在 $f(\mu) \leqslant \nu$ 和 $f(\mu) = \nu$ 时，称 f 为从 μ 到 ν 的**弱同构**或**同构**，或称 μ 弱同构或同构于 ν，分别记作 $\mu \overset{f}{\cong} \nu$，或 $\mu \overset{f}{=} \nu$.

在以上各记号中，f 也可以略去.

定义 14 设 $R \in L^{A \times A}$，如果 R 是 A 上的等价 L 模糊关系，且 $R \in L(A \otimes A)$，这里 $A \otimes A$ 表示泛代数 A 与其自身的直积代数，则称 R 为 A 上的 **L 模糊同余关系**，简称为 LF 同余关系.

定义 15 设 A 是 Ω 型泛代数，R 是 A 上的 L 模糊同余关系，按定义 11，A 关于 R 有一商集 $\frac{A}{R} = \{R_x : x \in X\}$.在此商集上定义 Ω 型运算如下：

1° $\forall \omega \in \Omega(0)$，令 R_ω 为 $\frac{A}{R}$ 的特定常元，且 $\frac{A}{R}$ 只有这些特定常元.

2° 设 $\omega \in \Omega(n)$, $x_1, x_2, \cdots, x_n \in A$, $n = 1, 2, \cdots$，令

$$\omega(R_{x_1}, R_{x_2}, \cdots, R_{x_n}) = R_\omega(x_1, x_2, \cdots, x_n).$$

可证上述 n 元运算的定义是合理的.称带有以上常元与运算的 Ω 型代数 $\frac{A}{R}$ 为 A 关于 R 的商代数.设 $\mu \in L(A)$，定义 $\frac{A}{R}$ 的 L 模糊子集 $\frac{\mu}{R}$ 如下：

$$\frac{\mu}{R}(R_x) = \bigvee_{y \in A} (\mu \wedge R_x)(y) \quad (x \in A),$$

其中 R_x 是 x 陪集.可证 $\frac{\mu}{R} \in L(A/R)$.$\frac{\mu}{R}$ 称为 μ 关于 R 的商.

定理 7 设 A, B 为 Ω 型泛代数，$\mu \in L(A)$, $\nu \in L(B)$, R 是 B 上的同余关系，则

1° 当 $\mu \sim \nu$ 时，存在 A 上的同余关系 Q 使 $\frac{\mu}{Q} \simeq \frac{\nu}{R}$；

2° 当$\mu \approx \nu$时，存在A上的同余关系Q使

$$\frac{\mu}{Q} \cong \frac{\nu}{R}.$$

2.2.3 L模糊子群

设G是群，e是其单位．则G是Ω型代数，这里$\Omega(0)=\{e\}$，$\Omega(1)=\{-1\}$，$\Omega(2)=\{\cdot\}$，$\Omega(n)=\varnothing(n\geqslant 3)$，所以2.2.2小节中关于$L$模糊子代数的理论均适用于$G$．把定义12用于$G$，可得如下定义．

定义16 设G是群，e是其单位，$\mu\in L^G$，如果

1° $\mu(e)=1$，

2° $\mu(x^{-1})\geqslant\mu(x) \qquad (x\in G)$，

3° $\mu(xy)\geqslant\mu(x)\wedge\mu(y) \qquad (x,y\in G)$，

则称μ为G的**L模糊子群**，简称为LF子群．其全体记作$L(G)$．

由定理6知，μ是G的L模糊子群，当且仅当$\forall\lambda\in L$，μ_λ是G的分明子群．

命题3 设$\mu\in L(G)$．$\forall\lambda\in L$，令$\mu^\lambda=e_l\vee\lambda\mu_\lambda$，则$\mu^\lambda\in L(G)$，称$\mu^\lambda$为$\mu$的**$\lambda$水平子群**．$\mu$可通过其$\lambda$水平子群而分解，即

$$\mu=\vee\{\mu^\lambda:\lambda\in\mu(G)\}, \tag{2-13}$$

其中$\mu(G)=\{\mu(g):g\in G\}$．当$\mu(G)=L$时，由(2-13)式得

$$\mu=\vee\{\lambda\mu_\lambda:\lambda\in L\}.$$

这是一般L模糊集的分解定理．

定义17 设$\mu\in L(G)$，$g\in G$，定义$\nu\in L^G$为

$$\nu(x)=\mu(gxg^{-1}) \qquad (x\in G). \tag{2-14}$$

则易证$\nu\in L(G)$．设$\mu,\nu\in L(G)$，若存在$g\in G$使(2-14)式成立，则称μ与ν为G的**LF共轭子群**．

定义18 设$\mu\in L(G)$，若对于G中任意二元x与y，均有

$$\mu(xy)=\mu(yx),$$

则称μ为G的**正规L模糊子群**．其全体记作$NL(G)$．

易证$\mu\in NL(G)$，当且仅当$\forall\lambda\in L$，μ_λ是G的分明正规子群．又$NL(G)$关于任意交运算封闭，即G的任意多个正规L模糊子群的交仍为G的正规L模糊子群．$NL(G)$关于定向并运算也是封闭的，即G的一族正规L模糊子群若为定向L模糊集族，则其并仍为G的正规L模糊子群．

定义19 设$\mu\in NL(G)$，$\forall x\in G$，定义G的L模糊子集$x\mu$及相应的乘法运算如下：

$$x\mu(t)=\mu(x^{-1}t) \quad (t\in G),$$

$$(x\mu\cdot y\mu)(t)=(xy)\mu(t)=\mu(y^{-1}x^{-1}t) \quad (t\in G).$$

称$\{x\mu:x\in G\}$是以$e\mu$为单位的群，叫做**G关于μ的商群**，记作G/μ．

命题4 设$\mu\in NL(G)$，令$\mu^*=\mu_1=\{x\in G:\mu(x)=1\}$，则$\mu^*$是$G$的分明正规子群，且

$$\frac{G}{\mu} \cong \frac{G}{\mu^*}.$$

命题5　设 G,H 是群，$f:G\twoheadrightarrow H$ 是满同态，若 $\mu\in NL(G)$，则 $f(\mu)\in NL(H)$；设 $f:G\to H$ 是同态，若 $\nu\in NL(H)$，则 $f^{-1}(\nu)\in NL(G)$.

正规子群的概念还可推广到 G 的 L 模糊子群之间.

定义20　设 $\mu,\nu\in L(G)$，且 $\mu\leqslant\nu$. 如果

$$\mu(xyx^{-1})\geqslant\mu(y)\wedge\nu(x)\quad(x,y\in G),\tag{2-15}$$

则称 μ 为 ν **的正规子群**.

当 $\nu=G$，即 $\nu=1_G$ 时，(2-15) 式成为

$$\mu(xyx^{-1})\geqslant\mu(y)\quad 或\ \mu(xy)\geqslant\mu(yx)\quad(x,y\in G).$$

由此得

$$\mu(xy)=\mu(yx)\quad(x,y\in G).$$

可见定义20是定义18的推广.

2.2.4　L 模糊子环

L 模糊子环也是 L 模糊子代数概念的特例.

定义21　设 R 是环，$\mu\in L^R$. 若

1° $\mu(0)=1$,

2° $\mu(-x)\geqslant\mu(x)$,

3° $\mu(x+y)\geqslant\mu(x)\wedge\mu(y)$, $\mu(xy)\geqslant\mu(x)\wedge\mu(y)$,

则称 μ 为 R **的 L 模糊子环**，简称为 LF 子环. 其全体记作 $L(R)$. 设 $\mu\in L(R)$，若当 $x,y\in R$ 时，

$$\mu(xy)\geqslant\mu(y)\qquad(\mu(xy)\geqslant\mu(x)),$$

则称 μ 为 R **的 L 模糊左理想**(**L 模糊右理想**). 其全体记作 $LI_l(R)(LI_r(R))$. 称

$$LI(R)=LI_l(R)\cap LI_r(R)$$

中的 L 模糊子环 μ 为 R **的 L 模糊双边理想**，简称为 IF 理想.

命题6　设 $\mu\in L^R$，则 μ 是 R 的极大 L 模糊左(右，双边)理想，即 μ 是 $LI_l(R)\setminus R(LI_r(R)\setminus R, LI(R)\setminus R)$ 中的极大元，当且仅当 μ_1 是 R 的分明极大左(右、双边)理想，且

$$\mu=\mu_1\vee\{\lambda\},$$

其中 λ 是 $L\setminus\{1\}$ 中的极大元.

由此可见，当 $L\setminus\{1\}$ 中没有极大元. 比如，当 $L=[0,1]$ 时，R 没有极大 L 模糊左(右，双边)理想，因此引入如下定义.

定义22　设 $\mu\in LI_l(R)(LI_r(R),LI(R))$，当 μ_1 是 R 的分明极大左(右，双边)理想时，称 μ 为 R **的广义极大 L 模糊左(右，双边)理想**.

命题7　设 R 是带单位的交换环，$\mu\in L^R$，则 μ 是 R 的广义极大 L 模糊理想，当且仅当 μ 可表示为

$$\mu=\mu_1\vee\{\lambda\},$$

其中 μ_1 是 R 的分明极大理想，且 $\lambda\in L\setminus\{1\}$.

定义 23 设 $\xi \in LI(R)$. ξ 叫做既约的,若当 $\mu,\nu \in LI(R)$,且 $\xi = \mu \vee \nu$ 时,有 $\xi = \mu$ 或 $\xi = \nu$. ξ 叫素的,若当 $\mu,\nu \in LI(R)$,且 $\mu \cdot \nu \leqslant \xi$ 时,$\mu \leqslant \xi$ 或 $\nu \leqslant \xi$. 这里 $\mu \cdot \nu$ 的定义是

$$(\mu \cdot \nu)(x) = \vee \{\mu(y) \wedge \nu(z) : yz = x\} \qquad (x \in R).$$

命题 8 设 R 是带单位的交换环,$\xi \in LI(R)$,则 ξ 是 R 的既约 L 模糊真理想,当且仅当 ξ_1 是 R 的分明既约真理想,且 $|\xi(R)| = 2$,即 ξ 是只取两个值的 L 模糊集.

命题 9 设 ξ 是 R 的既约 L 模糊理想,环同态 $f: R \to S$ 是满射,则当条件"$f(x) = f(y) \Rightarrow \xi(x) = \xi(y), x, y \in R$"成立时,$f(\xi)$ 是 S 的既约 L 模糊理想.

命题 10 设 $\xi \in L^R$,则 ξ 是 R 的 L 模糊素理想,当且仅当 ξ 可写为 $\xi = \xi \vee [c]$,这里 c 是 L 中的素元,即当 $a, b \in L$,且 $a \wedge b \leqslant c$ 时,有 $a \leqslant c$ 或 $b \leqslant c$.

定义 24 设 $\mu \in L^R$,定义 $\sqrt{\mu} \in L^R$ 如下:

$$\sqrt{\mu}(x) = \bigvee_{n=1}^{\infty} \mu(x^n) \quad (x \in R).$$

称 $\sqrt{\mu}$ 为 μ 的根.

易证 $\mu \leqslant \sqrt{\mu}$, $\sqrt{\sqrt{\mu}} = \sqrt{\mu}$.

命题 11 设 R 是带单位的交换环,$\nu \in LI(R)$,则 $\sqrt{\nu} \in LI(R)$,且

$$\sqrt{\nu} = \{\mu : \mu \in LI(R) \text{ 且 } \exists n \in N \text{ 使 } \mu^n \leqslant \nu\},$$

其中

$$\mu^n(u) = \vee \{\mu(x_1) \wedge \cdots \wedge \mu(x_n) : x_1 x_2 \cdots x_n = x\} \quad (x \in R).$$

2.2.5 L 模糊线性子空间

定义 25 设 V 是域 F 上的线性空间,$\mu \in L^V$. 如果

1° $\mu(0) = 1$,

2° $\mu(rx + sy) \geqslant \mu(x) \wedge \mu(y) \quad (r, s \in F, x, y \in V)$,

则称 μ 为 V 的 **L 模糊线性子空间**,简称为 LF 子空间,其全体记作 $L(V)$.

设 A 是任一代数,μ 是 A 的 L 模糊子集或分明子集. 由定理 6,存在由 μ 生成的 A 的子代数 $\langle\mu\rangle$.

命题 12 设 V 是域 F 上的线性空间,$\mu \in L^V$,则

$$< \text{supp}\,\mu > = \text{supp} < \mu > .$$

其中 $< \mu >$ 是由 μ 生成的 V 的 L 模糊线性子空间.

定义 26 设 $\mu \in L^V, x \in V$. 定义 $\mu \setminus x$ 为

$$(\mu \setminus x)(y) = \begin{cases} \mu(y) & (y \neq x), \\ 0 & (y = x). \end{cases}$$

称 μ 为**自由 L 模糊集**,若 $\forall x \in \text{supp}\,\mu$, $< \mu \setminus x > (x) = 0$.

命题 13 设 $\mu \in L^V$,则 μ 是自由的,当且仅当 $\text{supp}\,\mu$ 在 V 中线性无关.

定义 27 设 $\xi \in L^V$. 若 ξ 是自由的,且 $< \xi > = \mu$,则称 ξ 为 μ 的基.

命题 14 设 $\mu \in L(V)$,则

1° 若 μ 只取有限多个值,则 μ 有一个基,

2° 若 $\mu = \langle \nu \rangle$, ν 只取有限多个值,且 L 是全序集,则 μ 有一个基,

3° 若 V 是有限维空间,且 L 是全序集,则 μ 有一个基.

由于篇幅所限,以上关于 L 模糊代数只能作初步的介绍.有兴趣的读者可参看文献[5].

2.3　L 模糊拓扑空间

设 X 是非空集,L 是模糊格,与 L 模糊代数不同,在建立 X 上的 L 模糊拓扑空间理论时,并不要求 X 上有拓扑结构.又在讨论 L 模糊测度与积分以及 L 模糊代数时,往往只要求 L 的完备性,连分配性都不常用到.但在讨论 L 模糊拓扑时却一定要求 L 是模糊格,即具有逆序对合对应的分子格.这时 L 是完全分配的,从而 L 有一特殊的子集 M,$0 \bar{\in} M$,M 中的元叫**分子**,L 的每个元都可表示为分子之并,且对于 L 的元 A 与 B 以及分子 a,当 $a \leqslant A \vee B$ 时,有 $a \leqslant A$ 或 $a \leqslant B$.本节中恒假定 L 是模糊格,M 是其全部分子之集.这时 L^X 作为 $|X|$ 个模糊格的直积也是模糊格,其分子之集为

$$L^X(M) = \{x_\lambda : x \in X, \lambda \in M\},$$

即 L^X 中的分子之集由一切高度是 L 中分子的 F 点构成.又设 $f: L^X \to L^Y$ 为序同态,则可证 f 把 L^X 中的分子映射为 L^Y 中的分子.

定义 28　设 X 是非空集,L 是模糊格,$\delta \subseteq L^X$.如果

1° $0_X, 1_X \in \delta$,

2° δ 对有限交运算与任意并运算封闭,

则称 δ 为 X 上的 **L 模糊拓扑**,简称为 **LF 拓扑**.称(L^X, δ) 为 **L 模糊拓扑空间**,简称为 **LF 拓扑空间**.δ 中的 L 模糊集叫开元.

设 $\beta \subseteq \delta$,如果 δ 中每个元都可表示为 β 中若干元之并,则称 β 为 δ 的基.设 $\nu \subseteq \delta$,如果 ν 中元的有限交的全体构成 δ 的基,则称 ν 为 δ 的**子基**.令 $\eta = \{A' : A \in \delta\}$,称 η 中的 L 模糊集为闭元.

设 $A \in L$,令

$$\begin{cases} A^\circ = \bigvee \{B \in \delta : B \leqslant A\}; \\ A^- = \bigwedge \{C \in \eta : A \leqslant C\}. \end{cases} \tag{2-16}$$

分别称 A° 与 A^- 为 A 的**内部**与**闭包**.当 $L = [0,1]$ 时,(L^X, δ) 将简记为(X, δ),并称为**模糊拓扑空间**,简称为 **F 拓扑空间**.这时 L^X 中的全部分子之集就是 X 上的全部 F 点之集.

定义 29　设 $\delta = \{0_X, 1_X\}$,则(L^X, δ) 是 L 模糊拓扑空间,称为 X 上**平凡 L 模糊拓扑空间**.设 $\delta = L^X$,则(L^X, δ)是 L 模糊拓扑空间,称为 X 上的**离散 L 模糊拓扑空间**.设$\{\delta_i : i \in I\}$ 是 X 上一族 L 模糊拓扑,令 $\delta = \bigcap_{i \in I} \delta_i$,则 δ 仍是 X 上的 L 模糊拓扑.可见 X 上的全体 L 模糊拓扑之集 $\mathscr{T}$ 按包含序构成一完备格.又若 $\forall \lambda \in L$, $[\lambda] \in \delta$,即 δ 包括 X 上的一切常值 L 模糊集,则称(L^X, δ) 为**满层 L 模糊拓扑空间**.

定义 30 设$(X,\mathscr{U})$是分明拓扑空间，$A\in L^X$. 如果$\forall a\in L$，$\{x\in X:A(x)\leqslant a\}$是$(X,\mathscr{U})$中的闭集，则称$A$为$(X,\mathscr{U})$上的**$L$值下半连续函数**. 以$\omega_L(\mathscr{U})$记$(X,\mathscr{U})$上的全部$L$值下半连续函数之集，则可证$\omega_L(\mathscr{U})$是$L^X$上的$L$模糊拓扑，称$(L^X,\omega_L(\mathscr{U}))$为由$(X,\mathscr{U})$诱导出的$L$模糊拓扑空间，简称为**诱导空间**. 反过来，设(L^X,δ)为任一L模糊拓扑空间，$\forall A\in\delta$，$\forall a\in L$，令$\xi_a(A)=\{x\in X:A(x)\leqslant a\}$，令$\mathscr{S}(\delta)=\{(\xi_a(A))':a\in L,A\in\delta\}$，则以$\mathscr{S}(\delta)$为子基，可在$X$上生成一个分明拓扑，记作$\iota_L(\delta)$，叫做$\delta$的**截拓扑**.

可以证明，设$\mathscr{U}$是X上的分明拓扑，则$\iota_L(\omega_L(\mathscr{U}))=\mathscr{U}$. 又若$(L^X,\delta)$是诱导空间，则$\omega_L(\iota_L(\delta))=\delta$.

定理 8 设(L^X,δ)为L模糊拓扑空间，$A\in L^X$，则对A以任何顺序施行取内部运算、取闭包运算或取伪补运算任意多次，最多可得出X上的14个不同的L模糊集来，称为**14集定理**.

证明概要 利用德·摩根对偶律先证明$A'^{-\prime}=A^{\circ}$，再证明$A^{-\circ-\circ}=A^{-\circ}$，$A^{\circ-\circ-}=A^{\circ-}$，然后即可证明定理8成立.

定义 31 设(L^X,δ)是L模糊拓扑空间，$x_\lambda\in L^X(M)$. 如果$P\in\eta$，且$x_\lambda\not\leqslant P$，则称P为x_λ的**闭远域**. 设$Q\in L^X$，如果x_λ有闭远域P使$P\geqslant Q$，则称Q为x_λ的**远域**. x_λ的全体远域和闭远域之集，分别记作$\eta(x_\lambda)$和$\eta^-(x_\lambda)$. 易证，$\eta(x_\lambda)$是下集，即当$P\in\eta(x_\lambda)$，且$Q\leqslant P$时，$Q\in\eta(x_\lambda)$. 并且当$P,Q\in\eta(x_\lambda)$时，$P\vee Q\in\eta(x_\lambda)$. 又$1_X\,\overline{\in}\,\eta(x_\lambda)$，所以$\eta(x_\lambda)$是$L^X$中的理想. $\eta^-(x_\lambda)$是其理想基.

定义 32 设(L^X,δ)是L模糊拓扑空间，$A\in L^X$，$x_\lambda\in L^X(M)$，如果$\forall P\in\eta(x_\lambda)$，$A\not\leqslant P$，则称$x_\lambda$为$A$的**附着点**. 设$D$是定向集，则称映射$S:D\to L^X(M)$为$L^X$中的**分子网**，记作$S=\{S(n),n\in D\}$. 如果$\forall P\in\eta(x_\lambda)$，$S$最终不在$P$中，即有$n_0\in D$，使当$n\geqslant n_0$时，$S(n)\,\overline{\in}\,P$，则称$S$**收敛于$x_\lambda$**，或$x_\lambda$是$S$的**极限点**，记作$S\to x_\lambda$. 如果$\forall P\in\eta(x_\lambda)$，$S$经常不在$P$中，即$\forall n_0\in D$，有$n\geqslant n_0$，使$S(n)\,\overline{\in}\,P$，则称$S$**聚于$x_\lambda$**，或$x_\lambda$是$S$的**聚点**，记作$S\infty x_\lambda$. 下面令$\lim S=\vee\{x_\lambda\in L^X(M):S\to x_\lambda\}$，$AdS=\vee\{x_\lambda\in L^X(M):S\infty x_\lambda\}$.

定理 9 设(L^X,δ)是L模糊拓扑空间，$A\in L^X$，$a\in L^X(M)$，则下列条件等价：

1° $a\in A^-$.

2° a是A的附着点.

3° A中有分子网收敛于a.

4° A中有分子网聚于a.

定义 33 设(L^X,δ)与(L^Y,ξ)都是L模糊拓扑空间，$f:L^X\to L^Y$是序同态. 如果$\forall B\in\xi$，$f^{-1}(B)\in\delta$，则称f为**连续序同态**. 设$a\in L^X(M)$，如果对于L^X中每个收敛于a的分子网S，$f(S)$作为L^Y中的分子网收敛于分子$f(a)$，则称**f在分子a处连续**. 如果存在一一对应$f:X\to Y$，使f诱导的L值扎德型函数$f:L^X\to L^Y$以及其逆$f^{-1}:L^Y\to L^X$都连续，则称(L^X,δ)与(L^Y,ξ)为**弱同胚**，记作

$$(L^X,\delta)\overset{f}{\simeq}(L^Y,\xi)\quad 或\quad (L^X,\xi)\simeq(L^Y,\xi).$$

这时称 f 为从 (L^X,δ) 到 (L^Y,ξ) 的弱同胚映射.

定理10 设 (L^X,δ) 与 (L^Y,ξ) 是 L 模糊拓扑空间,$f: L^X \to L^Y$ 是序同态,则下列条件等价:

1° f 连续.

2° (L^Y,ξ) 中的闭元在 f 下的原像是 (L^X,δ) 中的闭元.

3° ξ 有子基 ξ, $\forall B \in \xi, f^{-1}(B) \in \delta$.

4° $\forall a \in L^X(M)$, f 在 a 处连续.

5° $\forall A \in L^X, f(A^-) \subseteq (f(A))^-$.

6° $\forall B \in L^Y, (f^{-1}(B))^- \subseteq f^{-1}(B^-)$.

7° $\forall B \in L^Y, f^{-1}(B^\circ) \subseteq (f^{-1}(B))^\circ$.

定义34 设 (L^X,δ) 是 L 模糊拓扑空间,$a \in L^X(M)$,如果 $\eta_0 \subseteq \eta(a)$,且 $\forall P \in \eta(a)$,有 $Q \in \eta_0$,使 $Q \geqslant P$,则称 η_0 为 $\eta(a)$ 的**远域基**.

如果 $\forall a \in L^X(M)$,a 有可数的远域基,则称 (L^X,δ) 为满足**第一可数公理**的 L 模糊拓扑空间,或 C_{I} 空间.

如果 δ 有可数基,则称 (L^X,δ) 为满足**第二可数公理**的 L 模糊拓扑空间,或 C_{II} 空间.

可以证明,C_{II} 空间一定是 C_{I} 空间,但反之不真.

注1 也可像一般拓扑学中那样对分子 a 引入其邻域系的概念.但基于邻域去定义闭包与收敛将导致矛盾,特别是 C_{I} 空间中分子不必有可数的邻域基.设 $A \in \delta$,且 $a \bar{\in} A'$,则称 A 为 a 的**开重域**,这等价于 A' 是 a 的闭远域,所以也可用重域去刻画闭包与收敛等.L 上逆序对合对应的存在性是使用重域工具的先决条件.当 L 上没有逆序对合对应时,远域方法仍然有效,这时只须直接给出闭元族 η 就行,而重域方法就无法使用了.

命题15 设 $(X,\omega(\mathscr{U}))$ 是诱导空间,则 $(X,\omega(\mathscr{U}))$ 是 C_{I} 空间或 C_{II} 空间,当且仅当 $(X,\mathscr{U})$ 相应地是 C_{I} 空间或 C_{II} 空间.

在 C_{I} 空间中,只须使用特殊的分子网——分子序列就可以刻画闭包、收敛与连续性等概念.

定义35 设 (L^X,δ) 是 L 模糊拓扑空间,如果当 λ,μ 是 L 中的分子,且 $\lambda < \mu$ 时,$\forall x \in X$,有 $P_x \in \eta(x_\mu)$ 使 $x_\lambda \leqslant P$,则称 (L^X,δ) 为 T_{-1} 空间.

如果对于 $L^X(M)$ 中任两个不同分子 x_λ 与 y_μ,有 $P \in \eta(x_\lambda)$ 使 $y_\mu \leqslant P$,或有 $Q \in \eta(y_\mu)$ 使 $x_\lambda \leqslant Q$,则称 (L^X,δ) 为 T_0 空间.

如果对于 L 中每个分子 λ,当 $x,y \in X$,且 $x \neq y$ 时,有 $P \in \eta(x_\lambda)$ 使 $y_\lambda \leqslant P$,或有 $Q \in \eta(y_\lambda)$ 使 $x_\lambda \leqslant Q$,则称 (L^X,δ) 为次 T_0 空间.

如果 $L^X(M)$ 中每个分子都是闭的,则称 (L^X,δ) 为 T_1 空间.

如果对于 $L^X(M)$ 中任两个承点不同的分子 x_λ 与 y_μ,有 $P \in \eta(x_\lambda)$ 和 $Q \in \eta(y_\mu)$ 使 $P \vee Q = 1_X$,则称 (L^X,δ) 为 T_2 空间或**豪斯多夫空间**.

如果 (L^X,δ) 是 T_1 空间,且 $\forall x_\lambda \in L^X(M)$,以及包含 x_λ 的开元 U,有开元 V 满足 $x_\lambda \leqslant V \leqslant V^- \leqslant U$,则称 (L^X,δ) 为 T_3 空间.

如果(L^X,δ)是T_1空间，且对于任一闭元K，以及包含K的开元U，有开元V满足$K\leqslant V\leqslant V^-\leqslant U$，则称$(L^X,\delta)$为**$T_4$空间**.

L模糊拓扑空间中的分离性理论远较一般拓扑学为复杂，有兴趣的读者可参看文献[6]与[7]. 下面是L模糊拓扑空间中有关分离性的若干结果.

命题 16　设(L^X,δ)是L模糊拓扑空间.

1° 若(L^X,δ)是T_1空间，则它也是T_0空间，又T_0空间也是T_{-1}空间和次T_0空间.

2° (L^X,δ)是T_{-1}空间的充要条件为，$\forall x_\lambda\in L^X(M)$，$x_\lambda$是$x_\lambda^-$中的极大分子.

3° (L^X,δ)是豪斯多夫空间，当且仅当对于L^X中任一分子网S，当$S\to x_\lambda$与$S\to y_\mu$时，$x=y$.

定义 36　设$\{(L^{X_t},\delta_t):t\in T\}$是一族$L$模糊拓扑空间，$T\neq\varnothing$，令$X=\prod_{t\in T}X_t$，$\forall t\in T$，$P_t:L^X\to L_{X_t}$是射影映射(即由分明射影映射$P_t:X\to X_t$诱导出的$L$值扎德型函数)，则在$L^X$上，以

$$\gamma=\{P_t^{-1}(A_t):A_t\in\delta_t,t\in T\}\tag{2-17}$$

为子基所生成的拓扑δ称为各L模糊拓扑$\{\delta_t:t\in T\}$的**乘积拓扑**，(L^X,δ)称为各L模糊拓扑空间的**乘积空间**.

定理 11　设$\{(L^{X_t},\delta_t):t\in T\}$是一族$L$模糊拓扑空间，$T\neq\varnothing$，$X=\prod_{t\in T}X_t$，则在$L^X$上的乘积拓扑$\delta$是使每个射影映射$P_t:(L^X,\delta)\to(L^{X_t},\delta_t)$都连续的$L^X$上的最弱$L$模糊拓扑，即$\delta$是满足当$\forall t\in T$，$P_t:(L^X,\mu)\to(L^{X_t},\delta_t)$都连续时，$\delta\subseteq\mu$的$L$模糊拓扑.

定义 37　设(L^X,δ)是L模糊拓扑空间，L_1是另一模糊格，$f:L^X\to {L_1}^Y$是满序同态，令

$$\mu=\{B\in {L_1}^Y:f^{-1}(B)\in\delta\},\tag{2-18}$$

则μ是${L_1}^Y$上的拓扑，称$({L_1}^Y,\mu)$为(L^X,δ)关于f的商空间，f称为**商序同态**.

定理 12　设(L^X,δ)是L模糊拓扑空间，$f:L^X\to {L_1}^Y$是满序同态，则${L_1}^Y$上的商拓扑μ是使f连续的最强拓扑，即设$f:(L^X,\delta)\to({L_1}^Y,\xi)$连续，则$\xi\subseteq\mu$.

定义 38　设$\mathbf{R}$是实直线，L是模糊格，令

$\Sigma=\{\lambda\in L^{\mathbf{R}}$：当$t<0$时，$\lambda(t)=0$，当$t>1$时，$\lambda(t)=0$，$\lambda$是递减函数$\}$.

设$\lambda,\mu\in\Sigma$. 规定$\lambda\sim\mu$，当且仅当$\forall t\in\mathbf{R}$，

$$\lambda(t+)=\mu(t+),\quad \lambda(t-)=\mu(t-),$$

即λ与μ作为$\mathbf{R}$上的L值函数处处有相等的右、左极限，则$\sim$是Σ上的等价关系. $\forall\lambda\in\Sigma$，以$[\lambda]$记λ所在的等价类. 令$\Omega=\Sigma/\sim$，$\forall t\in\mathbf{R}$，定义L^X中的L模糊集l_t与r_t如下：

$$\forall[\lambda]\in\Omega,\ l_t([\lambda])=(\lambda(t-))',$$
$$r_t([\lambda])=\lambda(t+).$$

令$\mathscr{L}=\{l_t:t\in\mathbf{R}\}$，$\mathscr{R}=\{r_t:t\in\mathbf{R}\}$，设$\theta$是$L^\Omega$上以$\mathscr{L}\cup\mathscr{R}$为子基生成的$L$模糊

拓扑,以 θ^* 记由 $\theta \cup \iota_L(\theta)$ 生成的 L 模糊拓扑,这里 $\iota_L(\theta)$ 是 θ 的截拓扑,则称 (L^Ω, ι^*) 为 $H(\lambda)$ 单位区间,记作 $\tilde{I}(L)$. 这里 H 表示赫顿(Hutton)单位区间,λ 表示 Ω 上的劳森(H. B. Lawson)拓扑. 可证,θ^* 是将赫顿单位区间的拓扑 θ 经过加入 $\lambda(\Omega)$ 而细化后得出的.[①]

定义 39 设 (L^X, δ) 是 L 模糊拓扑空间,如果 $\forall U \in \delta$,存在 L 模糊集族 $\{w_t: t \in T\}$,使得 $\bigvee_{t\in T} w_t = U$,且 $\forall t \in T$,存在连续的 L 值扎德型函数 $f_t:(L^X, \delta) \to \tilde{I}(L)$,使

$$w_t \leqslant f_t^{-1}(l_1') \leqslant f_t^{-1}(r_0) \leqslant U,$$

则称 (L^X, δ) 为 $H(\lambda)$ 完全正则空间. 如果 (L^X, δ) 还是 T_1 的,则称其为 $T_{3\frac{1}{2}}$ 空间.

定理 13 $T_{3\frac{1}{2}}$ 空间的乘积空间是 $T_{3\frac{1}{2}}$ 空间.

定义 40 设 (L^X, δ) 是 L 模糊拓扑空间,$A, B \in L^X$,如果 $A^- \wedge B = A \wedge B^- = O_X$,则称 A 与 B 是隔离的. 设 $C \in L^X$,如果不存在异于 O_X 的隔离集 A 与 B,使 $C = A \vee B$,则称 C 是连通集. 当 1_X 是连通集时,称 (L^X, δ) 是连通的 L 模糊拓扑空间,简称为连通空间.

命题 17 设 (L^X, δ) 是 L 模糊拓扑空间.

1° (L^X, δ) 是连通空间,当且仅当不存在异于 O_X 的闭(开)集 A 与 B,使 $A \vee B = 1_X$, $A \wedge B = O_X$.

2° 设 A 是 (L^X, δ) 中的连通集,$A \leqslant B \leqslant A^-$,则 B 是连通集.

3° 设 $\{A_t: t \in T\}$ 是 (L^X, δ) 中一族连通集,且有 $s \in T$,使 $\forall t \in T \setminus \{s\}$,$A_s$ 与 A_t 都不是隔离的,则 $\bigvee_{t\in T} A_t$ 是连通集.

4° (L^X, δ) 中每个连通集都包含于某极大连通集,称后者为**连通分支**.

5° (L^X, δ) 中的连通分支是闭元,不同的连通分支互不相交,且各连通分支的并等于 1_X.

6° 连通集在连续序同态下的像是连通集.

7° 一族连通空间的乘积空间是连通空间.

定理 14 (樊畿(K. Fan) 定理) 设 (L^X, δ) 是 L 模糊拓扑空间,$A \in L^X$,以 $M^*(A)$ 表示 A 中全体分子之集,$\forall x \in M^*(A)$,以 $\eta(x)$ 表示 x 的全部远域之集,那么,A 连通的充要条件为,对于每个映射

$$P: M^*(A) \to \bigcup \{\eta(x): x \in M^*(A)\}, P(x) \in \eta(x), x \in M^*(A),$$

以及 A 中任意两个分子 a 与 b,在 A 中,可找出有限多个分子 $x_0, x_1, \cdots, x_n$,使 $x_0 = a, x_n = b$,且

$$A \nleqslant P(x_i) \vee P(x_{i+1}). \quad (i = 0, 1, \cdots, n-1).$$

例 2 当 L 中的最大元 1 是分子时,$H(\lambda)$ 单位区间 $\tilde{I}(L)$ 是连通的.

---

① 王国俊,徐罗山. 内蕴拓扑与 Hutton 单位区间的细致化. 中国科学,1992,A 辑(7):705 ~ 712

定义 41 设 L 是完备格,$a \in L$,$B \subseteq L$,如果 $\vee B = a$,且当 $C \subseteq L$,$\vee C \geqslant a$ 时,$\forall x \in B$,有 $y \in C$,使 $x \leqslant y$,则称 B 为 a 的**极小集**.

可证,a 的全部极小集之并仍为 a 的极小集,记作 $\beta(a)$. 又 $\beta(a)$ 中的全体分子也组成 a 的极小集,叫做 a 的**标准极小集**,记作 $\beta^*(a)$. 可证,L 是分子格(即完备的完全分配格)的充要条件为:$\forall a \in L$,a 有一极小集,从而 $\beta^*(a)$ 存在.

定义 42 设 (L^X,δ) 是 L 模糊拓扑空间,$S = \{s(n), n \in D\}$ 是 L^X 中的分子网,以 $V(S(n))$ 表示分子 $S(n)$ 的高度,例如,当 $S(n) = x_\lambda$ 时,$V(S(n)) = \lambda$. 令 $V(S) = \{V(S(n)), n \in D\}$,这时 $V(S)$ 是 L 中的分子网,称其为 S 的**值网**.

设 a 是 L 中的分子,如果对于任一 $r \in \beta^*(a)$,S 的值网最终大于或等于 r,则称分子网 S 为 **a 网**.

设 $A \in L^X$,如果 A 中任一 a 网在 A 中有高度等于 a 的聚点,则称 A 为良紧集. 当 1_X 是良紧集时,称 (L^X,δ) 为良紧空间.

例 3 $H(\lambda)$ 单位区间 $\bar{I}(L)$ 是良紧空间.

定理 15 设 (L^X,δ) 是 L 模糊拓扑空间,A 是良紧集.

1° 设 C 是 (L^X,δ) 中的闭集,则 $A \wedge C$ 是良紧集.

2° 设 $f:(L^X,\delta) \to (L^Y,\xi)$ 是连续的 L 值扎德型函数,则 $f(A)$ 是 (L^Y,ξ) 中的良紧集.

3° 当 $L = [0,1]$ 时,A 作为 X 上的函数可取得最大值.

4° 如果 (L^X,δ) 是诱导空间,$\delta = \omega_L(\mathscr{U})$,则 (L^X,δ) 是良紧空间,当且仅当分明拓扑空间 $(X,\mathscr{U})$ 是紧空间.

定理 16 (吉洪诺夫(Tychonoff) 定理) 一族良紧空间的乘积空积是良紧空间.

关于紧化理论与仿紧空间理论,以及度量化等问题请参看文献[6]与[7]. 可以证明,本节中讨论的各种性质都是弱拓扑不变性质,即在弱同胚映射下被保持的性质.

2.4 L 模糊拓扑线性空间

2.4.1 L 模糊线性代数

设 V 是域 F 上的线性空间. 在 2.2.5 小节中已讨论了 V 的 L 模糊子空间的概念. 本节讨论 V 上的一些特殊的 L 模糊子集,这里 F 为实数域或复数域.

定义 43 设 V 是域 F 上的线性空间,$A, B \in L^V$,$r \in F$,规定 $A + B \in L^V$ 和 $rA \in L^V$ 如下:

$$(A+B)(x) = \vee \{A(y) \wedge B(z): y + z = x\},$$

$$(rA)(x) = A\left(\frac{x}{r}\right) \quad (r \neq 0),$$

$$(0A)(x) = \begin{cases} \sup\{A(t): t \in v\} & (x = \theta), \\ 0 & (x \neq \theta), \end{cases}$$

其中 $x \in V$,θ 表示域 V 的零元. 有

1° 若 $\forall r \in (0,1)$, $rA + (1-r)A \leqslant A$,则称 A 为**凸 L 模糊集**;

2° 若当 $|r| \leqslant 1$ 时, $rA \leqslant A$,则称 A 为**平衡 L 模糊集**;

3° 若 $V = \bigvee \{rA: r > 0\}$,则称 A 为**吸收 L 模糊集**,称平衡的凸 L 模糊集为**绝对凸 L 模糊集**;

4° 包含 A 的一切凸 L 模糊的交为凸 L 模糊集,叫做 A 的**凸包**,记作 co(A),包含 A 的一切平衡的 L 模糊集之交是平衡的 L 模糊集,叫做 A 的**平衡包**,记作 ba(A).

命题 18　设 V 是域 F 上的线性空间, $A \in L^V$,则有

1° A 是凸 L 模糊集,当且仅当 $\forall r \in [0,1]$,以及对于任意 $x, y \in V$,

$$A(rx + (1-r)y) \geqslant A(x) \wedge B(y),$$

当且仅当 $\forall \lambda \in L$, A_λ 是 V 中的分明凸集;

2° A 是平衡的 L 模糊集,当且仅当 $A(rx) \geqslant A(x)$ 对于 V 满足 $|r| \leqslant 1$ 的 r 均成立,当且仅当 $\forall \lambda \in L$, A_λ 是 V 中的分明平衡集;

3° A 是吸收的 L 模糊集,当且仅当 $\forall x \in V$,存在 $\varepsilon > 0$,使当 $|t| < \varepsilon$ 时, $tx_\mu \,\tilde{\in}\, A$,当且仅当 $\forall \lambda \in L$, A_λ 是 V 中的分明吸收集;

4° $\mathrm{co}(A) = \bigvee \{\sum_{i=1}^{n} r_i A_i : r_i \geqslant 0, \sum_{i=1}^{n} r_i = 1, n = 1,2,\cdots\}$;

5° $\mathrm{ba}(A) = \bigvee \{rA: |r| \leqslant 1\}$.

2.4.2　模糊拓扑线性空间

本节中设 L 为单位区间,从而 L 模糊集就成为模糊集.

定义 44　设 X 是域 F 上的线性空间, F 为实数域或复数域, (X,δ) 是模糊拓扑空间, $f: X^2 \to X$ 与 $g: F \times X \to X$ 是分明映射,分别由 $f(x,y) = x + y$ 与 $g(r,x) = rx (x, y \in X, r \in F)$ 定义.以 $\mathscr{J}$ 记 F 上的分明拓扑,则 $(F, \mathscr{J})$ 也可看做模糊拓扑空间.如果 f 与 g 诱导的扎德型函数 $f: (X,\delta)^2 \to (X,\delta)$ 与 $g: (F,\mathscr{J}) \times (X,\delta) \to (X,\delta)$ 都连续,则称 (X,δ) 为**模糊拓扑线性空间**,简称为 **F 拓扑线性空间**.

下面域 F 常省略而不提及.

定义 45　设 X 是线性空间, $A, B \in [0,1]^X$,规定 $A \oplus B \in [0,1]^X$ 如下:

$$(A \oplus B)(x) = \wedge \{A(y) \vee B(z): y + z = x\} \qquad (x \in X).$$

1° 若 $\forall r \in (0,1)$, $B \leqslant rB \oplus (1-r)B$,则称 B 为余凸集.

2° 若当 $0 < |r| \leqslant 1$ 时, $B \leqslant rB$,且 $B(\theta) \leqslant \inf B$,则称 B 为余平衡集.

3° 若 $\wedge \{rB: r > 0\} = O_X$,则称 B 为余吸收集.设 $\lambda \in [0,1]$,若 $\{x \in X: B(x) < \lambda\}$ 是 X 中的分明吸收集,则称 B 为 **λ 下吸收集**.

4° 包含于 B 的一切余凸集之并为余凸集,称为 B 的余凸核,记作 cok(B).

5° 包含于 B 的一切余平衡集之并为余平衡集,称为 B 的余平衡核,记作 bak(B).

命题 19　设 X 是线性空间, (X,δ) 是拓扑空间,则定义 44 中的扎德型函数 f 连续的充要条件为,对于任意二模糊点 x_λ 与 y_μ,设 $P \in \eta(x_\lambda + y_\mu)$,则有 $Q \in \eta(x_\lambda)$ 与 $R \in \eta(y_\mu)$,使 $P \leqslant Q \oplus R$.定义 44 中的扎德型函数 g 连续的充要条件为,对于

域 F 中任一常数 r 和任一模糊点 x_λ，设 $P \in \eta(rx_\lambda)$，则有 $Q \in \eta(x_\lambda)$ 及 $\delta > 0$，当 $|s - r| < \delta, s \neq 0$ 时，$P \leqslant sQ$，且 $P(\theta) \leqslant \inf Q$.

命题 20 设(X,δ)是模糊拓扑线性空间，$x_0 \in X, t_0 \in F, f_0$与$g_0$分别是由分明映射 $f_0(x) = x_0 + x$ 与 $g_0(x) = t_0(x \in X)$ 诱导的扎德型函数，则

$$(X,\delta) \overset{f_0}{\simeq} (X,\delta),$$
$$(X,\delta) \overset{g_0}{\simeq} (X,\delta).$$

即 f_0 与 g_0 都是从(X,δ)到自身的弱同胚映射.

命题 21 设(X,δ)是模糊拓扑线性空间，则有

1° $P \in \eta(\theta_\lambda)$，当且仅当 $x + P \in \eta(x_\lambda)$，

2° $P \in \eta(x_\lambda)$，当且仅当 $rP \in \eta((rx)_\lambda), r \neq 0$.

以上两个命题在一定意义上反映了 F 拓扑线性空间的均匀性.

命题 22 设(X,δ)是模糊拓扑线性空间，B 是余平衡集，则 B 的核 B° 是余平衡集.

定理 17 设(X,δ)是模糊拓扑线性空间，则 θ_λ 的远域 P 称为正规的远域，若 $\forall x \in X, P(\theta) \leqslant P(x)$. 设 $\forall \lambda \in (0,1], \eta^\circ(\theta_\lambda)$ 是 θ_λ 的正规远域基，$P \in \eta^\circ(\theta_\lambda)$，则

1° 有 $Q \in \eta^\circ(\theta_\lambda)$，使 $P \leqslant Q \oplus Q$;

2° 有 $Q \in \eta^\circ(\theta_\lambda)$，使当 $0 < |t| \leqslant 1$ 时，$P \leqslant tQ$，且 $P(\theta) \leqslant \inf Q$;

3° $\forall \varepsilon \in (0,\lambda)$，有 $Q \in \eta^\circ(\theta_\lambda)$，使 $Q \geqslant P \vee [\lambda - \varepsilon]$;

4° 设 $x \in X, P(x) > \mu > 0$，则有 $Q \in \eta^\circ(\theta_{1-\mu})$，使 $P \leqslant x_1 \oplus Q$;

5° 设 $x \in X, \lambda - \varepsilon > \mu > 0$，则有 $Q \in \eta^\circ(\theta_{1-\mu})$，使$[\lambda - \varepsilon] \leqslant x_1 \oplus Q$;

6° $\forall x \in X$，有 $\alpha > 0$，使 $\alpha P(x) > 1 - \lambda$.

定理 18 设(X,δ)是模糊拓扑线性空间，则 $\forall \lambda \in (0,1], \forall P \in \eta^-(\theta_\lambda)$，定理 17 中的条件 1° ~ 6° 成立.

定理 19 设(X,δ)是模糊拓扑线性空间，则 $\forall \lambda \in (0,1], \theta_\lambda$ 有一正规的、余平衡的、且 λ 下吸收集的远域基.

定理 20 设(X,δ)是模糊拓扑线性空间，则下列条件等价：

1° (X,δ) 是 T_2 空间；

2° $\forall \lambda \in (0,1]$，模糊点 θ_λ 是闭集；

3° $\forall \lambda \in (0,1], \forall x \in X$，当 $x \neq \theta$ 时，有 $P \in \eta(\theta_\lambda)$，使 $P(x) = 1$;

4° $\forall x \in X$，当 $x \neq \theta$ 时，有 $P \in \eta(\theta_1)$，使 $P(x) = 1$.

关于有界性与局部凸性等理论可参看文献[1]与[3].

3 应用举例

模糊数学理论已被应用到许多方面，其中既包括模糊数学中新方法的应用，也包括模糊集思想的应用，也就是在经典数学方法中把某类数据适当地模糊化以后

所得方法的应用.由于篇幅所限,这里仅介绍模糊聚类理论和模糊推理理论.读者从中可见模糊数学应用之一斑.

3.1　模糊聚类

模糊聚类方法由于应用背景的不同而有多种不同的方法.本节中介绍基于模糊等价关系的聚类方法和基于目标函数的聚类方法.

3.1.1　基于模糊等价关系的聚类方法

设 $X=\{x_1,x_2,\cdots,x_n\}$ 是一组有限多个对象.所谓聚类,也就是分类,而分类当然要依据各对象的某种性质去分.如,X 是 n 个人的集合,所考虑的性质是年龄,那么只要预先定好几个年龄段,是很容易将 X 按年龄分类的.这时 X 也可按其他性质进行分类,如,根据性别分类(最多只有两类了)、根据体重分类(如对举重运动员所作的分类那样)、根据职业分类、根据学历分类,等等.一般的聚类分析理论假定被考虑的性质不是一个而是一组(s 个性质),即每个对象 x_i 对应着按一组性质去考察的结果$(x_{i1},x_{i2},\cdots,x_{is})$,或者干脆就把 x_i 等同于其特征向量:

$$x_i=(x_{i1},x_{i2},\cdots,x_{is})\quad(i=1,2,\cdots,n).\tag{3-1}$$

现在要根据(3-1)式对 X 进行分类.对于每个固定的 $j(1\leqslant j\leqslant s)$,$x_{1j},x_{2j},\cdots,x_{nj}$ 表示 n 个对象的同一类量化了的性质,它们都是数值,自然容易比较其间的亲疏远近,而要将 s 个性质通盘考虑来决定任二对象间的相似程度,则按实际问题的性质不同而有许多不同的算法.以 r_{ij} 记 x_i 与 x_j 间的相似度,为方便起见,希望 r_{ij} 能具有如下特点:

$$\begin{aligned}&1^\circ\ 0\leqslant r_{ij}\leqslant 1\quad(i,j=1,2,\cdots,n),\\&2^\circ\ r_{ii}=1\qquad\quad(i=1,2,\cdots,n),\\&3^\circ\ r_{ij}=r_{ji}\qquad(i,j=1,2,\cdots,n).\end{aligned}\tag{3-2}$$

使 r_{ij} 满足以上三条件的方法非常多,如:

(1) **夹角余弦法**　夹角余弦法是令

$$r_{ij}=\frac{\sum\limits_{k=1}^{s}x_{ik}x_{jk}}{\left(\left(\sum\limits_{k=1}^{s}x_{ik}^2\right)\left(\sum\limits_{k=1}^{s}x_{jk}^2\right)\right)^{1/2}};$$

(2) **算术平均与取小法**　算术平均与取小法是令

$$r_{ij}=\frac{\sum\limits_{k=1}^{s}x_{ik}\wedge x_{jk}}{\frac{1}{2}\sum\limits_{k=1}^{s}(x_{ik}+x_{jk})};$$

(3) **绝对值指数法**　绝对值指数法是令

$$r_{ij} = \exp\left(-\sum_{k=1}^{s} | x_{ik} - x_{jk} |\right),$$

等等.这里不再列出其他算法,因为读者可以根据实际问题的性质,按(3-2)式中的原则去试探更加合适的能表征各对象之间相似程度的算式.特别是当对象的一组 s 个特征不宜被平等看待时,还可利用加权的办法改进 r_{ij} 的计算.

至此已得出一个 $n \times n$ 矩阵 $\boldsymbol{R} = [r_{ij}]$. $\boldsymbol{R}$ 显然是 X 上的一个自反的且对称的模糊关系,在 X 有限的情形,称其为模糊矩阵. $\boldsymbol{R}$ 一般不是传递的.只须求出 $\boldsymbol{R}$ 的等价闭包 $\boldsymbol{R}^*$,然后就可通过求 $\boldsymbol{R}^*$ 的 $\lambda(\lambda \in (0,1])$ 截集,而将 X 进行分类了. λ 越大,分的类就越细,各类中对象的相似程度也越大.反之,λ 越小,分的类就越粗糙,各类中对象的相似程度也越小.

例 设 $X = \{x_1, x_2, \cdots, x_6\}$,

$$x_1 = (3,5,-2,6), x_2 = (4,3,-3,7),$$
$$x_3 = (0,1,-1,2), x_4 = (9,2,-2,3),$$
$$x_5 = (9,3,-4,3), x_6 = (2,4,-4,8).$$

按算术平均与取小法计算各对象间的相似度.

解 $r_{12} = 0.78$, $r_{13} = 0.33$, $r_{14} = 0.50$, $r_{15} = 0.43$, $r_{16} = 0.73$,

$r_{23} = 0.35$, $r_{24} = 0.52$, $r_{25} = 0.55$, $r_{26} = 0.76$,

$r_{34} = 0.83$, $r_{35} = 0.70$, $r_{36} = 0.09$,

$r_{45} = 0.87$, $r_{46} = 0.27$,

$r_{56} = 0.38$,

所以,得到一个 6×6 自反的,且对称的模糊矩阵 $\boldsymbol{R}$ 如下:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.78 & 0.33 & 0.50 & 0.43 & 0.73 \\ 0.78 & 1 & 0.35 & 0.52 & 0.55 & 0.76 \\ 0.33 & 0.35 & 1 & 0.83 & 0.70 & 0.09 \\ 0.50 & 0.52 & 0.83 & 1 & 0.87 & 0.27 \\ 0.43 & 0.55 & 0.70 & 0.87 & 1 & 0.38 \\ 0.73 & 0.76 & 0.09 & 0.27 & 0.38 & 1 \end{bmatrix}.$$

$\boldsymbol{R}$ 的等价闭包 $\boldsymbol{R}^* = \bigvee_{n=1}^{\infty} \boldsymbol{R}^n$.事实上只须计算 $\boldsymbol{R}^2, \boldsymbol{R}^4, \boldsymbol{R}^8, \cdots$,直到算出 $\boldsymbol{R}^k = \boldsymbol{R}^{2k}$ 为止,$\boldsymbol{R}^k$ 就是所求的 $\boldsymbol{R}^*$.按此方法可求得 $\boldsymbol{R}^* = \boldsymbol{R}^4$,即

$$\boldsymbol{R}^* = \begin{bmatrix} 1 & 0.78 & 0.55 & 0.55 & 0.55 & 0.76 \\ 0.78 & 1 & 0.55 & 0.55 & 0.55 & 0.76 \\ 0.55 & 0.55 & 1 & 0.87 & 0.87 & 0.55 \\ 0.55 & 0.55 & 0.87 & 1 & 0.87 & 0.55 \\ 0.55 & 0.55 & 0.87 & 0.87 & 1 & 0.55 \\ 0.76 & 0.76 & 0.55 & 0.55 & 0.55 & 1 \end{bmatrix}.$$

容易验证 $\boldsymbol{R}^*$ 确为模糊等价矩阵.现在作 $\boldsymbol{R}^*$ 的 λ 截集以对 X 进行分类.若取 $\lambda > 0.87$,如 $\lambda = 0.9$,或 $\lambda = 1$,则得 $\boldsymbol{R}^*_\lambda = \boldsymbol{I}_X$ 为 X 上的单位矩阵.这时分类过细,每个

类只有一个对象.若取 $\lambda = 0.87$,则得

$$R_\lambda^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

这时 X 被分成4类:

$$\{x_1\},\{x_2\},\{x_3,x_4,x_5\},\{x_6\}.$$

若取 $\lambda = 0.78$,则得

$$R_\lambda^* = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

这时 X 被分成3类:

$$\{x_1,x_2\},\{x_3,x_4,x_5\},\{x_6\}.$$

若取 $\lambda = 0.76$,则得

$$R_\lambda^* = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

这时 X 被分成2类:

$$\{x_1,x_2,x_6\},\{x_3,x_4,x_5\}.$$

若取 $\lambda \leqslant 0.55$,则整个 X 就成为一类了.

3.1.2 基于目标函数的聚类方法

基于模糊等价关系的聚类方法只适用于 X 中的对象不太多的情形.当 X 中对象个数 n 很大时,一般采用基于目标函数的聚类方法.

设等待分类的对象集为 $X = \{x_1,x_2,\cdots,x_n\}$,其中每个 x_i 都有 s 个特征,即 $x_i \in \mathbf{R}^s$.先假定已知 X 应该被分成几类,比如,分成 c 类,$1 < c < n$.这里在 $c = 1$ 和 $c = n$ 两个极端情形下,X 分别被看做是一个类,或被分成 n 个类(一个对象算一个类),这当然是无价值的情形,因而不予考虑.设分类数 c 已经确定,则有如下的**模糊 c 均值聚类算法**,或简称为 **Fc 聚类算法**.其基本思想是:假定 $Y = \{y_1,y_2,\cdots,y_c\} \subseteq \mathbf{R}^s$ 是理想的 c 个聚类中心,即 $\mathbf{R}^s$ 中的 c 个点,这里的 $y_j(j = 1,2,\cdots,c)$ 不一

定属于 X. 当然希望各个类聚集得很紧密，这可通过要求各类中的对象到相应的聚类中心的距离平方和最小来实现. 令 d_{ij} 表示 $d(x_i, y_j)$，$1 \leqslant i \leqslant n$，$1 \leqslant j \leqslant c$，$d$ 是 $\mathbf{R}^s$ 中的欧氏距离. 以 $[y_j]$ 记 X 中以 y_j 为聚集中心的那些 x_i 组成的类. 令

$$u_{ij} = \begin{cases} 1 & (\text{当 } x_i \in [y_j]), \\ 0 & (\text{否则}), \end{cases}$$

则

$$\sum_{j=1}^{c} u_{ij} = 1 \qquad (i = 1, 2, \cdots, n). \tag{3-3}$$

令 U 为 $n \times c$ 阶的 0-1 矩阵 $[u_{ij}]$，令 $V = [y_1, y_2, \cdots, y_c]$.

注意：每个 y_j 都是 $\mathbf{R}^s$ 中一个点，是一个 s 维向量，所以 V 也可看做一个 $s \times c$ 阶矩阵. 令

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} d_{ij}^2, \tag{3-4}$$

如果 $\{y_1, y_2, \cdots, y_c\}$ 是理想的聚类中心，则(3-4)式应当取最小值. 称 $J(U, V)$ 为目标函数.

当待分类的对象数目很大时，往往一个对象 x_i 既可被分到 $[y_j]$ 中去，又似乎可被分到 $[y_k]$ $(k \neq j)$ 中去. 这时(3-3)式中的 u_{ij} 非 1 即 0，不能反映这种亦此亦彼的情况. 所以可将 u_{ij} 改为 $X \times Y$ 上的模糊集，满足条件

$$\begin{cases} \sum_{j=1}^{c} u_{ij} = 1, \\ 0 < \sum_{i=1}^{n} u_{ij} < n. \end{cases} \tag{3-5}$$

其中，$i = 1, 2, \cdots, n$；$j = 1, 2, \cdots, c$.

令

$$J_m(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m d_{ij}^2. \tag{3-6}$$

这里 $m > 1$ 是固定的常数. 所谓模糊 c 聚类算法就是要求目标函数 $J_m(U, V)$ 的最小值以及与此最小值相应的聚类中心和各对象属于各中心的程度 u_{ij}，$i = 1, 2, \cdots, n$，$j = 1, 2, \cdots, c$. 具体算法如下：

在 $\mathbf{R}^s$ 中任意固定 c 个聚类中心 $V^{(0)} = [y_1^{(0)}, y_2^{(0)}, \cdots, y_c^{(0)}]$. 预先给定一个计算精度 $\varepsilon > 0$. 令 $d_{ij}(0) = d(x_i, y_j^{(0)})$. 对于一个固定的 i，$1 \leqslant i \leqslant n$，若有 r，使 $d_{ir}(0) = 0$，则令

$$u_{ir}(0) = 1, u_{ij}(0) = 0 \quad (j \neq r).$$

若各 $d_{ir}(0)$ 均不为 0，则令

$$u_{ij}(0) = \frac{1}{\sum_{r=1}^{c} \left(\frac{d_{ij}(0)}{d_{ir}(0)} \right)^{2/(m-1)}}.$$

然后计算 $V^{(1)}$ 如下：

$$y_j^{(1)} = \frac{\sum_{i=1}^{n} u_{ij}^m(0) x_i}{\sum_{i=1}^{n} u_{ij}^m(0)} \quad (j = 1,2,\cdots,c). \tag{3-7}$$

注意,上式中 x_i 与 $y_j^{(1)}$ 都是 s 维向量.

设 $V^{(k)} = [y_1^{(k)}, y_2^{(k)}, \cdots, y_c^{(k)}]$, $d_{ij}(k)$, $u_{ij}(k)$ 已求出,则可按(3-7)式的方式,求 $V^{(k+1)}$:

$$y_j^{(k+1)} = \frac{\sum_{i=1}^{n} u_{ij}^m(k) x_i}{\sum_{i=1}^{n} u_{ij}^m(k)} \quad (j = 1,2,\cdots,c, k = 1,2,\cdots).$$

直到求出 $V^{(k)}$ 使 $\| V^{(k)} - V^{(k+1)} \| < \varepsilon$ 为止,$\| \cdot \|$ 是 $s \times c$ 维欧氏空间中的模.可以证明,按上述算法可求得一局部最优解 (U^*, V^*),满足

$$J_m(U^*, V^*) \leqslant J_m(U, V^*), \text{对于一切 } U \text{ 均成立},$$

$$J_m(U^*, V^*) \leqslant J_m(U^*, V), \text{对于一切 } V \text{ 均成立}.$$

以上假定了已知 X 应当被分为 c 类.究竟应分成几类,当然可以结合实际情况就不同的 c 进行计算与比较而确定,这样计算量很大.但如何简捷地确定最佳分类数是个很复杂的问题,至今未圆满解决.现在得到的结果为

1° 根据实际发现,一般分类数 c 应不大于 $2\ln n$.

2° c 的确定是后验的,而非先决的,即针对若干满足 $c \leqslant 2\ln n$ 的 c,按以上算法求出相应的 $U = [u_{ij}]$ 来,然后比较 $f(U,c)$ 的值,按 f 的不同取法,这个值越大或越小,表示分类数 c 越合理.如

(i) 取 f 为划分系数 $F(U,c)$:求 F 的最大值.

$$F(U,c) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^2.$$

(ii) 取 f 为划分熵 $H(U,c)$:求 H 的最小值.

$$H(U,c) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} \ln u_{ij}.$$

(iii) 取 f 为 $F(U,c) - P(U,c)$:求 $F - P$ 的最小值.

$$P(U,c) = \frac{1}{c} \sum_{j=1}^{c} \left(\frac{\sum_{i=1}^{n} u_{ij}^2}{\sum_{i=1}^{n} u_{ij}} \right).$$

如此等等①.

---

① 范九伦.模糊聚类新算法与聚类有效性问题研究:[博士学位论文].西安:西安电子科技大学,1998.

3.2 模糊推理

模糊推理是模糊控制的理论基础，在工业生产领域中有重要的应用.如今普遍流行的是由扎德于1975年提出的CRI方法.这一方法已经成功地应用于许多方面.但从理论基础上看，尚有不足之处.

本节着重介绍三I算法，它是CRI算法的改进.为简便计，以下用"(k)式"表示"(3-k)式".

3.2.1 模糊推理的基本思想

一个系统的输出与预定的标准之间总是会有误差的，以 e 记此误差.这个误差又是随着时间而变化的，以 $\dot{e}$ 记误差变化率.所谓控制就是要根据 e 与 $\dot{e}$（或许还需要 e 的多阶导数）对系统的输入进行调整.经典控制理论中这个调整量是 e 和 $\dot{e}$ 的函数 $f(e,\dot{e})$.在某些情况下这个 f 是不易建立或不需要建立的，取而代之的是已知一组专家经验，即已知当误差为 e_i，且变化率为 $\dot{e}_i$ 时，应当采取的调整量为 c_i，$i=1,2,\cdots,n$，要在这一组典型情况的基础上，针对随时测得的 e^* 与 $\dot{e}^*$ 计算出相应的控制量 c^* 来.

模糊控制的基本原理是：

第一步，把 $e_i,\dot{e}_i,c_i,e^*$ 与 $\dot{e}^*$ 分别**模糊化**为 X,Y,Z,X 与 Y 上的模糊集 A_i,B_i,C_i,A^* 与 B^*.

第二步，列出算式：

$$\begin{array}{lc} \text{已知} & A_1 \text{ 且 } B_1 \longrightarrow C_1 \\ & \cdots\cdots\cdots \\ & A_n \text{ 且 } B_n \longrightarrow C_n \\ \underline{\text{且给定}} & \underline{A^* \text{ 且 } B^* \qquad\quad} \\ \text{求} & \qquad\qquad C^* \end{array} \tag{3-8}$$

称为**模糊推理**.

第三步，将 C^* **去模糊**(defuzzify)后，就得到最终的数值控制量 c.由于条件 A_i 和 B_i 可用乘积 $X\times Y$ 上的一个模糊集去取代，且 n 条规则可以通过**聚合**(aggregate)而成为一条超规则，或者可以分别使用这 n 条规则单独推理后，将所得 n 个中间结果以某种方式合成为最终的 C^*，所以模糊推理可归结为以下最基本的形式：

$$\begin{array}{lc} \text{已知} & A \longrightarrow B \\ \underline{\text{且给定}} & \underline{A^* \qquad\quad} \\ \text{求} & \qquad B^* \end{array} \tag{3-9}$$

这里 A,A^* 是 X 上的模糊集，B,B^* 是 Y 上的模糊集.

求解(3-9)式的传统方式,是扎德于1975年提出的CRI算法①.扎德提出了一种蕴涵算子 $R_Z:[0,1]^2 \to [0,1]$,即

$$R_Z(a,b) = a' \vee (a \wedge b). \tag{3-10}$$

其中 a' 表示 $1-a$, $\vee$ 与 $\wedge$ 分别表示取上、下确界运算.CRI算法的第一步是利用蕴涵算子 R_Z 把已知条件 $A \to B$ 转化为 $X \times Y$ 上的一个模糊关系 $R(x,y)$,即

$$R(x,y) = R_Z(A(x),B(y)) \quad ((x,y) \in X \times Y). \tag{3-11}$$

然后在第二步用 A^* 与 R 复合得出 B^*,即 $B^* = A^* \circ R$.这里复合运算的具体表达式为

$$\begin{aligned} B^*(y) &= \sup_{x\in X}[A^*(x) \wedge R(x,y)] \\ &= \sup_{x\in X}[A^*(x) \wedge R_Z(A(x),B(y))] \quad (y \in Y). \end{aligned} \tag{3-12}$$

这样就求出了最终答案 B^*.

3.2.2 模糊推理的三I算法

20多年来虽然有不少学者提出过不同于CRI算法的模糊推理算法,CRI算法自身也有各种变形与发展(参看文献[8] ~ [13]),但都采取了复合运算这一步.那么为什么要采取复合运算呢?似乎至今没有一个令人满意的理由.因为蕴涵算子恰好是与推理相配的,所以利用某种蕴涵运算把 $A \to B$ 转化为模糊关系这一想法是好的.可惜CRI算法没有沿此思路走下去.事实上,由(3-9)式求出的 B^* 可以看做是在已知 $A \to B$ 的前提之下,由 A^* 推导出来的,这里自然应当把 $A^* \to B^*$ 考虑进来,即同样利用蕴涵算子(比如 R_Z)把它转化为另一个模糊关系

$$R_1(x,y) = R_Z(A^*(x),B^*(y))((x,y) \in X \times Y).$$

这里对于每一对固定的 (x,y), $R_1(x,y)$ 表示了 $A^* \to B^*$ 的真实程度.自然希望这种真实程度越大越好.但 $A^* \to B^*$ 又是在 $A \to B$ 的前提之下提出的,所以应当要求

$$(A(x) \to B(y)) \to (A^*(x) \to B^*(y)) \tag{3-13}$$

的值越大越好.

注意,(3-13)式中有三重蕴涵关系.三I算法正是基于这种分析提出的.

以下分别用 $\mathscr{F}(X)$, $\mathscr{F}(Y)$, … 表示 X, Y, … 上的模糊集的全体.当(3-13)式中的 $A, A^* \in \mathscr{F}(X)$,以及 $B \in \mathscr{F}(Y)$ 都已知时, $\mathscr{F}(Y)$ 中使(3-13)式取得最大值的 B^* 显然是有的,比如,由 $R_Z(a,b)$ 关于 b 递增知,在 Y 上取常值1的 $\mathscr{F}(Y)$ 中的最大集 1_Y 就使(3-13)式取最大值.如果采用其他的蕴涵算子,如卢卡谢维奇(J. Lukasiewicz)蕴涵算子 R_L:

$$a \to b = R_L(a,b) = (a' + b) \wedge 1, \tag{3-14}$$

或哥德尔蕴涵算子 α:

---

① Zadeh L A. The concept of linguistic variable and its application to approximate reasoning. Inform Sci, 1975, (B): 199 ~ 249.

$$a \to b = \alpha(a,b) = \begin{cases} 1 & (a \leqslant b), \\ b & (a > b), \end{cases} \tag{3-15}$$

或根思 - 瑞舍(Gaines-Rescher)的蕴涵算子 R_{GR}：

$$a \to b = R_{GR}(a,b) = \begin{cases} 1 & (a \leqslant b), \\ 0 & (a > b), \end{cases} \tag{3-16}$$

等,则由 $a \leqslant b$ 时(a,b)在这些算子作用下的值都等于1可知,在(3-13)式中令 $B^* = 1_Y$,则(3-13)式的值达到最大值1.但这个1_Y不是所需要的.事实上,这个1_Y代表了恒真命题,无论怎么改变(3-13)式中的A,B与A^*,1_Y,总使(3-13)式达到最大值.换句话说,这时(3-13)式所以取得最大值并非建立在前提 $A \to B$,以及给定的A^*的基础上.所以应当寻求$\mathscr{F}(Y)$中能使(3-13)式取最大值的最小可能模糊集B^*,这种B^*才是恰好在已知$A \to B$时由A^*推出的那个模糊集.这是三I算法的另一个基本思想,即有下面的原则.

三I原则 设X,Y是非空集,$A,A^* \in \mathscr{F}(X)$,$B \in \mathscr{F}(Y)$,则(3-9)式中的B^*是使(3-13)式对于一切$(x,y) \in X \times Y$,取得最大值的$\mathscr{F}(Y)$中的最小模糊集①.

根据以上原则,如果采用扎德的蕴涵算子R_Z,则有下面的定理1.

定理1 设X,Y是非空集,$A,A^* \in \mathscr{F}(X)$,$B \in \mathscr{F}(Y)$,则(3-9)式中B^*的算法为

$$B^*(y) = \sup_{\substack{x \in E_y \\ R_Z(A(x),B(y)) > \frac{1}{2}}} [A^*(x) \wedge R_Z(A(x),B(y))] \quad (y \in Y). \tag{3-17}$$

亦称为扎德型三I算法,其中

$$E_y = \{x \in X \mid (A^*(x))' < R_Z(A(x),B(y))\}.$$

注意,$R_Z(a,a) = 1$一般不成立,即$a \to a$不必等于1.这似乎是R_Z的不足之处.它还会导致还原性(见下节)不成立.卢卡谢维奇算子虽具有较多的好性质,但在传递性方面有缺陷.所以引入在若干方面具有较好性质的算子$\mathbf{R_0}$②：

$$R_0(a,b) = \begin{cases} 1 & (a \leqslant b), \\ a' \vee b & (a > b). \end{cases} \tag{3-18}$$

下面主要使用蕴涵算子R_0.

定理2 设X,Y是非空集,$A,A^* \in \mathscr{F}(X)$,$B \in \mathscr{F}(Y)$,则(3-9)式中B^*的算法为

$$B^*(y) = \sup_{x \in E_y} [A^*(x) \wedge R_0(A(x),B(y))] \quad (y \in Y). \tag{3-19}$$

亦称为$\mathbf{R_0}$型三I算法,其中

① 王国俊.模糊推理的全蕴涵三I算法.中国科学,1999,E辑(1):43～53

② 王国俊.修正的Kleene系统中的$\sum$-(α-重言式)理论.中国科学,1998,E辑(2):146

$$E_y = \{x \in X \mid (A^*(x))' < R_0(A(x), B(y))\} \qquad (y \in Y). \qquad (3\text{-}20)$$

将(3-17)式与(3-12)式相比较可见，用三I算法和用CRI算法求B^*时，取上确界运算后的表达式完全一样．但(3-17)式中x的变化范围比(3-12)式中小，因而，对于每个$y \in Y$，由(3-17)式求出的$B^*(y)$比由(3-12)式求出的$B^*(y)$要小．从而按最小性原则可知，(3-17)式的结果较(3-12)式为优．

3.2.3 关系再现算法

在上一小节已经看到，扎德的CRI算法除了在原理上有缺陷外，其计算结果也不是最优的．它的另一个不足是当$A^* = A$时，$B^* = B$一般不成立，即CRI算法不是还原算法，或按文献[12]的术语，CRI算法不是**关系再现算法**．本节讨论三I算法的还原性问题．首先将三I算法推广用于模糊拒取式(fuzzy modus tollens，简称FMT)的情形．FMT的一般形式为

$$\begin{array}{lr} \text{已知} & A \longrightarrow B \\ \underline{\text{且给定} \qquad\qquad B^*} \\ \text{求} & A^* \end{array} \qquad (3\text{-}21)$$

其中A, A^*与B, B^*仍分别是X与Y上的模糊集．有如下原则和定理．

三I(FMT)原则 设X, Y是非空集，$A \in \mathscr{F}(X)$，$B, B^* \in \mathscr{F}(Y)$，则(3-21)式中的$A^*$是使(3-13)式对于一切$(x,y) \in X \times Y$，取得最大值的$\mathscr{F}(X)$中的最大模糊集．

定理3（R_0型三I(FMT)算法） 设X, Y是非空集，$A \in \mathscr{F}(X)$，$B, B^* \in \mathscr{F}(Y)$，则(3-21)式中$A^*$的算法为

$$A^*(x) = \inf_{y \in E_x}[B^*(y) \vee R_0'(A(x), B(y))] \quad (x \in X). \qquad (3\text{-}22)$$

其中

$$E_x = \{y \in Y \mid B^*(y) < R_0(A(x), B(y))\} \quad (x \in X). \qquad (3\text{-}23)$$

在很弱的条件下，对于FMP和FMT而言，三I算法都是关系再现算法．

定理4 设A是X上的正规模糊集，即有$x_0 \in X$，使$A(x_0) = 1$，则R_0型三I算法是关系再现算法．即若(3-9)式中$A^* = A$，则$B^* = B$．

定理5 设B'是Y上的正规模糊集，即有$y_0 \in Y$，使$B(y_0) = 0$，则R_0型三I(FMT)算法是关系再现算法．即若(3-21)式中$B^* = B$，则$A^* = A$．

顺便指出，按CRI算法，FMT的运算结果为

$$A^*(x) = \sup_{y \in Y}[B^*(y) \wedge R_Z(A(x), B(y))]. \qquad (3\text{-}24)$$

(3-24)式与(3-22)式相去甚远．考虑一个很简单的情形，令$X = Y = [0,1]$，$A(x) = x(x \in X)$，$B(y) = B^*(y) = y(y \in Y)$，这时由(3-18)式，可求出

$$A^*(x) = x \vee x' = x \vee (1 - x),$$

并不还原为x，而且这时(3-13)式成为

$$(x \to y) \to (x \vee x' \to y). \qquad (3\text{-}25)$$

它一般都不取最大值，甚至当$x = y = 0$时，它的值等于0．可见CRI算法在FMT的情形是不可取的．

3.2.4 支持度理论

三I算法的另一个优点是可以被方便地推广为 α 三I算法. 事实上,要求(3-13)式对于一切 $(x,y)\in X\times Y$ 都等于1,可以理解为要求 $A\to B$ 对 $A^*\to B^*$ 全力支持. 设 $\alpha\in[0,1]$,上述要求可以一般化为要求

$$(A(x)\to B(y))\to(A^*(x)\to B^*(y))\geqslant\alpha \tag{3-26}$$

对于一切 (x,y) 都成立.

α 三I原则 设 X,Y 是非空集, $A,A^*\in\mathscr{F}(X)$, $B\in\mathscr{F}(Y)$,则(3-9)式的 α 解 B^* 是使(3-26)式对于一切 $(x,y)\in X\times Y$ 都成立的 $\mathscr{F}(Y)$ 中的最小模糊集.

相应地,定理2可以推广为下面定理.

定理6 设 X,Y 是非空集, $A,A^*\in\mathscr{F}(X)$, $B^*\in\mathscr{F}(Y)$,则(3-9)式的 α 解 B^* 的算法为

$$B^*(y)=\sup_{x\in E_y\cap K_y}[A^*(x)\wedge \mathrm{R}_0(A(x),B(y))]\wedge\alpha\quad(y\in Y). \tag{3-27}$$

亦称为 **R_0 型 α 三I算法**,其中 E_y 的含义同(3-20)式;

$$K_y=\{x\in X\mid A^*(x)\wedge \mathrm{R}_0(A(x),B(y))>\alpha'\}\quad(y\in Y). \tag{3-28}$$

当 $\alpha=1$ 时,(3-27)式中的 α 不起作用,且因

$$K_y=\mathrm{supp}[A^*(x)\wedge \mathrm{R}_0(A(x),B(y))],$$

$x\in E_y\cap K_y$ 也可用 $x\in E_y$ 代替,可见这时(3-27)式转化为(3-19)式. 所以定理6是定理2的一般化形式.

关于FMT的一般化,有如下原则和定理.

α 三I(FMT)原则 设 X,Y 是非空集, $A\in\mathscr{F}(X)$, $B,B^*\in\mathscr{F}(Y)$,则(3-21)式的 α 解 A^* 是使(3-26)式对于一切 $(x,y)\in X\times Y$ 都成立的 $\mathscr{F}(X)$ 中的最大模糊集.

定理7 设 X,Y 是非空集, $A\in\mathscr{F}(X)$, $B,B^*\in\mathscr{F}(Y)$,则(3-21)式的 α 解 A^* 的算法为

$$A^*(x)=\inf_{y\in E_x\cap K_x}[B^*(y)\vee \mathrm{R}_0'(A(x),B(y))]\vee\alpha'\quad(x\in X). \tag{3-29}$$

亦称为 R_0 型 α 三I(FMT)算法,其中 E_x 的含义同(3-23)式;

$$K_x=\{y\in Y\mid B^*(y)\vee \mathrm{R}_0'(A(x),B(y))<\alpha\}\quad(x\in X). \tag{3-30}$$

定理3是定理7当 $\alpha=1$ 时的特例.

支持度概念还可以一般化而用于同一论域上的任何模糊集之间.

定义1 设 X 是非空集, $A,B\in\mathscr{F}(X)$,则称

$$\alpha=\inf\{A(x)\to B(x)\mid x\in X\} \tag{3-31}$$

为 A 对 B 的**支持度**,记作 $\mathrm{sust}(A,B)=\alpha$.

定义1中的 $A(x)\to B(x)$ 可借助任何蕴涵算子去计算. 如,当使用算子 R_0 时,称相应的支持度为 R_0 型支持度等.

R_0型支持度具有一定意义下的传递性.

定理8 设X是非空集,$A,B,C\in\mathscr{F}(X)$. 如果R_0型支持度

$$\mathrm{sust}(A,B)=\alpha>\frac{1}{2},\quad \mathrm{sust}(B,C)=\beta>\frac{1}{2},$$

则R_0型支持度

$$\mathrm{sust}(A,C)\geqslant\alpha\wedge\beta.$$

注意,如果采用**卢卡谢维奇蕴涵算子**R_L,则相应的支持度不具有传递性,即定理8不成立,这是算子R_L的一个不足之处.

R_0型支持度具有如下性质:

定理9 设X是非空集,$A,B,A_i,B_i\in\mathscr{F}(X)(i\in I)$,则有

1° $\mathrm{sust}(\bigvee_{i\in I}A_i,B)=\bigwedge_{i\in I}\mathrm{sust}(A_i,B)$.

2° $\mathrm{sust}(A,\bigwedge_{i\in I}B_i)=\bigwedge_{i\in I}\mathrm{sust}(A,B_i)$.

设X,Y是非空集,$B,B_i\in\mathscr{F}(X),C,C_i\in\mathscr{F}(Y)(i\in I),A\in\mathscr{F}(X\times Y)$,则

3° $\mathrm{sust}(A,\bigvee_{i\in I}B_i\to C)=\bigwedge_{i\in I}\mathrm{sust}(A,B_i\to C)$,

4° $\mathrm{sust}(A,B\to\bigwedge_{i\in I}C_i)=\bigwedge_{i\in I}\mathrm{sust}(A,B\to C_i)$.

3.2.5 Σ-(α-重言式)理论

为适应模糊命题演算的需要,引入一种形式演绎系统$\mathscr{L}^*$. 这里仅涉及它的语义部分.

定义2 设$S=\{p_1,p_2,\cdots\}$,$F(S)$是由S生成的$(\neg,\vee,\to)$型自由代数,则称$F(S)$的元素为命题或公式,称S中的元素为原子命题或原子公式.

定义3 设映射$v:F(S)\to[0,1]$是$(\neg,\vee,\to)$型同态,这里$[0,1]$是R_0区间. 即当$a,b\in[0,1]$时,$\neg a=a',a\vee b=\max(a,b),a\to b=R_0(a,b)$,则称$v$为$F(S)$的一个**赋值**. $F(S)$的全体赋值之集记作$\bar{\Omega}$. 对于$F(S)$中的公式A,也称$v(A)$为A的赋值.

由$F(S)$是自由代数可知,下面的命题1成立.

命题1 设$v_0:S\to[0,1]$是任一映射,则存在$F(S)$的唯一赋值$v:F(S)\to[0,1]$,使v是v_0的扩张. 称v为由v_0生成的$F(S)$的赋值.

定义4 设$A\in F(S),\Sigma\subset\bar{\Omega},\alpha\in(0,1]$,如果对于每个$v\in\Sigma$,恒有$v(A)\geqslant\alpha$,则称$A$为**Σ-(α-重言式)**.

当$\Sigma=\Omega$,且$\alpha=1$时,Σ-(α-重言式)就是重言式.

Σ-(α-重言式)理论即**部分赋值理论**,只要适当选取Σ,就可将模糊推理纳入于多值逻辑框架之中. 事实上,可把(3-13)式抽象化为

$$P=(p_1\to p_2)\to(p_3\to p_4).\tag{3-32}$$

由于$p_i(i=1,2,3,4)$可以各自独立赋值,所以(3-32)式中的P不是重言式,也不是α重言式,但适当选取$\bar{\Omega}$的子集Σ后,就可使一般的R_0型α三I算法问题成为求Σ并使P成为Σ-(α-重言式)问题.

定义5 设X,Y是非空集,$A,A^*\in\mathscr{F}(X),B,B^*\in\mathscr{F}(Y),E=(A,B;A^*,$

B^*).定义 $v_0=v_0(E,x,y):S\to[0,1]$为

$$v_0(p_1)=A(x),v_0(p_2)=B(y),v_0(p_3)=A^*(x),$$
$$v_0(p_4)=B^*(y),v_0(p_n)=0\qquad(n=5,6,\cdots).\tag{3-33}$$

以 $v(E,x,y)$记由 v_0 生成的 $F(S)$的赋值.定义

$$\Sigma(E)=\{v(E,x,y)\mid(x,y)\in X\times Y\}.\tag{3-34}$$

利用定义 5 就可把模糊推理从语义角度纳入于逻辑框架之中.

定理 10 设 X,Y 是非空集,$A,A^*\in\mathscr{F}(X)$,$B\in\mathscr{F}(Y)$,则(3-9)式的 α 解就是 $\mathscr{F}(Y)$中使(3-32)式的 P 成为Σ-(α-重言式)的最小模糊集 B^*,这里 $\Sigma=\Sigma(E)$,$E=(A,B;A^*,B^*)$.又设 $A\in\mathscr{F}(X)$,$B,B^*\in\mathscr{F}(Y)$,则(3-21)式的 α 解就是 $\mathscr{F}(X)$中使(3-32)式的 P 成为Σ-(α-重言式)的最大模糊集 A^*,这里 $\Sigma=\Sigma(E)$,$E=(A,B;A^*,B^*)$.

3.2.6 多条规则的模糊推理

具有多条推理规则的模糊推理(3-8)式可以归结为问题(3-9)式.一般先利用乘积方法把(3-8)式化为

$$\begin{array}{ll}\text{已知} & A_1\longrightarrow B_1\\ & \cdots\cdots\cdots\cdots\\ & A_n\longrightarrow B_n\\ \underline{\text{且给定}} & \underline{A^*\qquad\qquad}\\ \text{求} & \qquad\quad B^*\end{array}\tag{3-35}$$

求解(3-35)式有两种常见的途径:

第一种途径是先分别利用 $A_i\to B_i$ 与 A^* 求得中间结果 C_i,$i=1,2,\cdots,n$,然后把这 n 个中间结果聚合为最终结果 B^*.布克利(Buckley)等称这种方法为 FITA (first infer then aggregate).

第二种途径是先把(3-35)式中的 n 条规则聚合为一条超规则 $A\to B$.然后利用给定的 A^* 求得 B^*.布克利等称此为 FATI(first aggregate then infer),并称 FATI 是不相容的,即当 A^* 等于某 A_i 时,B^* 不等于 B_i.但其证明是错误的[①].

事实上,在一定意义的聚合之下 FATI 与 FITA 都可以是相容的,而且在三 I 算法之下二者是等价的.

定义 6 设 X,Y 是非空集,$A_i,A_*\in\mathscr{F}(X)$,$B_i\in\mathscr{F}(Y)$,$i=1,2,\cdots,n$,则(3-35)式的 FITA 型解 $B^*=\bigvee_{i=1}^{n}C_i$,这里 C_i 为

$$\begin{array}{ll}\text{已知} & A_i\longrightarrow B_i\\ \underline{\text{且给定}} & \underline{A^*\qquad\qquad}\\ \text{求} & \qquad\quad C_i\end{array}$$

① Buckley J, Hayashi Y. Fuzzy input-output controllers ar universal approximators. Fuzzy Sets and Systems, 1993(58):273~278

的解，$C_i \in \mathscr{F}(Y), i=1,2,\cdots,n$.

定义 7 设 X, Y 是非空集，$A_i, A^* \in \mathscr{F}(X)$, $B_i \in \mathscr{F}(Y)$, $i=1,2,\cdots,n$，则(3-35)式的 FATI 型解 B^* 是 $\mathscr{F}(Y)$ 中使

$$R(x,y) \to (A^*(x) \to B^*(y)) = 1, \tag{3-36}$$

对于一切$(x,y) \in X \times Y$恒成立的最小模糊集，其中

$$R(x,y) = \bigvee_{i=1}^{n} R_0(A_i(x), B_i(y)). \tag{3-37}$$

由蕴涵算子 R_0 的性质容易证明下面的定理.

定理 11 FITA 算法与 FATI 算法是等价的，即当 A_i, B_i 与 A^* 给定时，由定义 6 与定义 7 算出的 B^* 相等.

求解(3-35)式的另一种方法是在 A^* 与各 $A_i, i=1,2,\cdots,n$，之间引进距离或贴近度等概念，然后给定阈值让 A^* 激活(3-35)式中的某一条或某几条规则，并与它们一起去计算 B^*①. 又关于(3-35)式中规则的聚合问题，也可以先分组聚合后再将各结果进行聚合②.

参考文献

1 吴从炘，马明. 模糊分析学基础. 北京：国防工业出版社，1991.
2 胡淑礼. 模糊数学及其应用. 成都：四川大学出版社，1994.
3 张文修，王国俊，刘旺金等. 模糊数学引论. 西安：西安交通大学出版社，1991.
4 Wang Z Y, Klir G J. Fuzzy measure theory. New York: Plenum Press, 1992.
5 Yandong Y, Mordeson J N, Cheng S C. Elements of L-Algebra. Omaha: Creighton University, 1994.
6 王国俊. L-fuzzy 拓扑空间论. 西安：陕西师范大学出版社，1988.
7 Liu Y M, Luo M K. Fuzzy topology. Singapore: World Scientific, 1997.
8 吴望名，模糊推理的原理和方法. 贵阳：贵州科技出版社，1994.
9 徐扬，乔全喜，陈超平等. 不确定性推理. 成都：西南交通大学出版社，1994.
10 张文修，梁怡. 不确定性推理原理. 西安：西安交通大学出版社，1996.
11 陈永义. 模糊控制技术及应用实例. 北京：北京师范大学出版社，1993.
12 张文修，梁广锡. 模糊控制与系统. 西安：西安交通大学出版社，1998.
13 汪培庄. 模糊集理论及其应用. 上海：上海科技出版社，1983.
14 王国俊. 非经典数理逻辑与近似推理. 北京：科学出版社，2000.

① Wang G J. Fuzzy continuous input-output controllers are universal approximators, Fuzzy Sets and Systems, 1998(97): 95 ~ 100

② 王国俊. 论袋映射及其结构. 模糊系统与数学，1996, 3(10): 1 ~ 11

索　引

使用说明：1. 本索引收录了本卷正文中用黑体排印的大部分术语。

2. 术语依第一字的读音按汉语拼音字母表顺序排列。如果拼音相同，根据音调，按阴平、阳平、上声、去声、轻声的次序排列。如果音和音调也相同，按总笔画数排列。

3. 以符号、数字或字母起首的术语，按符号、数字、拉丁字母、希腊字母的顺序，分别集中排在以汉字起首的术语前面，其中字母依首写字母按字母的顺序排列。

4. 以数学家译名为首的术语（例如，傅里叶变换），依译名按汉字的排法排列。

5. 术语后面的数字，表示该术语出现在本书中的页码。

以符号、数字起首的术语

以拉丁字母起首的术语

以希腊字母起首的术语

以汉字起首的术语

A

B

C

D

E

F

G

K

L

M

N

O

P

Q

T

Y

图书在版编目(CIP)数据

现代数学手册·经济数学卷/《现代数学手册》编纂委员会
武汉:华中科技大学出版社,2001年1月
ISBN 7-5609-2176-0

Ⅰ.现…
Ⅱ.现…
Ⅲ.①数学-手册 ②经济数学-手册
Ⅳ.O 1-62

现代数学手册·经济数学卷 《现代数学手册》编纂委员会

责任编辑:佟文珍,叶见欣,姜新祺,周芬娜 封面设计:刘 卉
责任校对:张 欣 责任监印:张正林

出版发行:华中科技大学出版社 武昌喻家山 邮编:430074 电话:(027)87545012
经销:新华书店湖北发行所

录排:湖北省新华印刷厂
印刷:湖北省新华印刷厂

开本:880×1230 1/32 印张:27.375 插页:6 字数:1 100 000
版次:2001年1月第1版 印次:2001年1月第1次印刷 印数:1—8 000
ISBN 7-5609-2176-0/O·209 定价:80.00元